**Diseases Prediction Application Using Machine Learning Models.**

Team Members: Pinki Sharma, Shruti Shah, and Umaima Khurshid Ahmad

**Abstract:**

The inability of a human being to process huge amounts of information, analyze it, and apply it to particular disease diagnosis can lead to medical diagnostic errors. In this study, we want to build a chatbot that patients can interact with and based on patients' entered responses it will then predict a disease. The study uses MCA, KModes, built-in cloud technologies (Azure Auto ML), Decision Tree, Gradient Boosting, Random Forest, and XGBoost models to explore the research question. MCA and KModes were performed, it showed the relationship among categorical variables and how they are clustered together. XGboost classifier was used as a final mode for our front-end application. The application was successfully predicting a users’ disease.

**Introduction:**

The major aim of this project is to create a platform to automate disease diagnosis. Keeping in mind that a system developed won’t take over the job of an actual human doctor but would assist the doctors to do their job easily. This area has huge potential, given the fact that we have doctors but their availability in certain areas is low. There are only 6,000 diseases, out of which there are 1,000 that are common. There are about 50,000 signs and symptoms of these diseases. There are 100,000 types of lab reports. Matching these to diagnose a disease can be done by any software very easily, which takes a doctor years to practice. There is recent research in China, that a robot passed the Medical Licensing Doctor Exam which potential doctors pass with years of studying.

There are some key reasons why we need disease diagnosis to be done by machines:

1. Every year 195,000 patients in the US die of medical diagnostic errors.
2. Inability of a human being to process huge amounts of information, analyze it, and apply it to a particular disease diagnosis. The cloning of an expert robot is cheap and fast compared to a doctor, where the cloning of expert human doctors is forbidden by law in many countries.

Our main aim from this project is to create a platform that can help doctors in disease diagnosis and to help individuals to cure disease in the early stage. Most models out in the market are trained with a large amount of data, for the project we aim to do it on the data we have.  We believe our system will help doctors to aid their decision-making abilities for disease diagnostics.

**Literature Review:**

Prior to analyzing the dataset, research was performed using three journal articles. The Journal Articles are named as follows:

1. Efficient Automated Disease Diagnosis Using Machine Learning Models [1]
2. Machine learning equipped web-based disease prediction and recommender system[2]
3. Disease prediction from various symptoms using machine learning [3]
4. Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning [4]
5. Building a Medical Chatbot using Support Vector Machine Learning Algorithm [5]
6. Doctor Chatbot: Heart Disease Prediction System [6]
7. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques [7]

The first article [1], Uses efficient machine learning algorithms to analyze a combined dataset of coronavirus, heart disease, and diabetes datasets merged to predict the risk of these diseases in an individual. An end-to-end process is used where people must enter their details in the mobile application and submit the data. In the proposed model, the data are entered into an android app. The analysis is then performed in a real-time database using a pre-trained machine learning model trained on the same dataset and deployed in firebase, and finally, the disease detection result is shown in the android app. Logistic regression is used to carry out computation for prediction. Extensive experimental results reveal that the proposed model outperforms the competitive machine learning models in terms of accuracy and F-measure by 1.4765% and 1.2782, respectively, for the COVID-19 dataset. The proposed model outperforms the competitive machine learning models in terms of accuracy and F-measure by 1.8274% and 1.7264, respectively, for the diabetes dataset. The proposed model outperforms the competitive machine learning models in terms of accuracy and F-measure by 1.7362% and 1.3821, respectively, for the heart disease dataset.

The second article [2], is about predicting disease based on symptoms given by the users, to help the healthcare department for easy access to the medical history of patients. The dataset used in the article is from survey data. KNN, Naive Bayes, Random Forest, and Ensemble models were initially used in an iterative approach. Ensemble classifiers performed better with an accuracy of 99.65% and precision and recall of 99.7818%. Web-based application(raahat) was used as a user interface. Django Framework acts as a medium to integrate the UI to the backend model in this paper.

The Third article [3], is again about disease prediction using Machine Learning Models. Dataset used was from Kaggle with 230 diseases and 1000 unique symptoms. Weighted KNN followed by Fine KNN gave the highest Accuracy in the prediction of diseases based on symptoms in comparison to Decision Trees and Naive Bayes as illustrated in the paper. Fine KNN is about taking just 1 neighbor and Weighted KNN is about giving more weights to closest neighbors, suggesting a pattern where the closer the data is to the symptoms, the higher the prediction Accuracy is for the symptoms.

The fourth article [4], is about disease prediction and recommending treatment through chatbot. The dataset consists of general information about symptoms and diseases. For building a chatbot, text processing was done using NLP and the model was built using K Nearest Neighbors(KNN), the data was split into a train and test set where 75% of the data was used for training and the rest 25% for testing. KNN identifies the symptoms from the interaction with the user, it maps the symptoms to a particular disease and provides a link where details about the treatment is visible. The chatbot was tested among people and a person having symptoms like dryness, cough and headache were analyzed based on the dataset and the medical chatbot correctly predicted the disease ‘cold fever’ for the given symptoms.

The fifth article [5], is about a medical chatbot using a support vector machine learning algorithm. The proposed model was to build a system that helps users to submit queries regarding their health. The model was trained under different datasets with 70% training and 30% testing set. The chatbot collects patients' basic personal details, symptoms, and other medical-related details. If the age of a patient is below 18, the chatbot will stop the conversation by telling them to inform their parents else the conversation continues. For simple testing, KNN and Naïve Bayes algorithms were used, but the Support Vector Machine (SVM) algorithm gave the most accurate results with an accuracy of 92%, so the model was built using SVM.

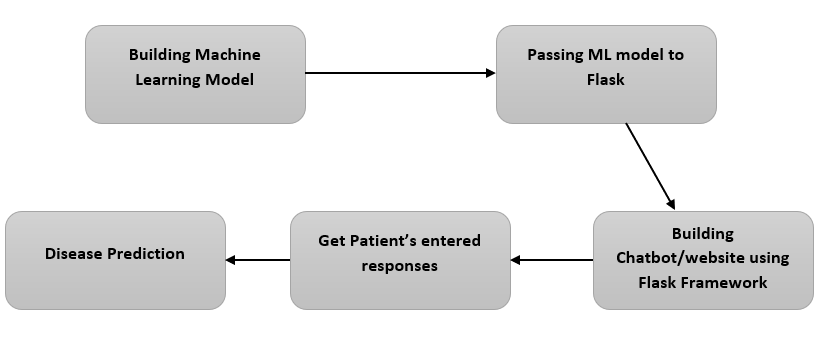
In the research paper [6], The data is not described where taken but generally told that heart disease prediction data. The tools used for creating chatbots are DialogFlow, Microsoft Bot Framework, Telegram Bot API, etc. Disease prediction can be done using various data mining algorithms along with their respective domain-specific datasets. The project is built to mostly execute on desktop Web Browsers, it is recommended to supply input in text format and not a voice so that input is wrong inputs – Giving an idea that the Speech to Text is not an optimized solution for their project. The researcher went ahead with SVM as it had the highest “accuracy” but it should have been used as the evaluation metric as precision and recall should also have been catered.

The research paper [7] aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. The proposed hybrid HRFLM approach was used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease.

**Dataset**

The research question was analyzed on the dataset of Disease Diagnosis hosted by Kaggle. The dataset contains 2124 rows and 13 variables out of which 7 are categorical variables and the rest 6 are continuous variables. The dataset contains some personal information of a patient such as age, weight, height, gender, BMI level as well as details like what symptoms they are having, what region they belong to, intensity, and severity of the symptom. The explanatory variables are symptom, age, weight, height, intensity, severity, and gender. Our target variable is a disease. The *Figure1(Appendix)* shows the snapshot of the data. After further feature engineering, selection, and handling imbalance our dataset consisted of 7138 rows.

**Methods**



*The Figure 2 :above shows project breakdown.*

Major analysis and changes were done on the dataset to investigate our research question experiments. The following analysis and changes were: i) dropped 5 variables unrelated to our investigations: 'Disease\_CUI', 'Symptom\_CUI', 'Season', 'Region' and ‘Severity’. ii) Symptoms and disease column data consists of numerical values in their name like coro162ry heart disease, so we replaced those 162 with “na”. iii) Combined different diseases so that the model has more similar data to train with. Approach - Used google to search 148 diseases,  came down to 60, used pandas data-wrangling approach to make a new dataset. We have combined similar symptoms like “chest pain” and “pain chest” as “chest pain”. iv) Then used MCA and KModes techniques in clean data to analyze categorical variables in order to understand how categorical variables are related. v) Categorical variables (Symptoms) were converted into one-hot encoding. vi) Handling the imbalanced dataset by oversampling the minority class. We further dropped the variable ‘BMI\_level’ from our analysis.

After considering the following changes above, we examined our model building on Decision Tree, Random Forest, Gradient Boosting. To dig deeper with our analysis, we also implemented Auto ML on Azure Cloud to verify what techniques an auto ML model takes; from the AutoML of Azure, we used the XGBoost hyperparameters on our local XGBoost model to reduce the model complexity. We tried to optimize the decision tree by giving parameters like max\_depth of range(1,10) and min\_samples split. Lastly, tried implementing PCA analysis in our dataset, because of the huge class label of categorical variables we didn't get proper analysis. PCA did not give an inherent relationship in our analysis of data.

After the data cleaning process, analyzing the data and building the model, we created a web-based application so the patient/user can connect with the system and the system can predict disease based on user response. We created a user interface (UI) page to connect with users. The UI page was created using HTML and all the model building processing was done using python language. We used “flask” to integrate UI with our backend (model). On the UI page we asked multiple questions like what your age is, gender, how intense is your pain, what symptom you have, how severe is your symptoms. Based on the response of these questions disease is predicted and the user will get a pop message saying, “ You may suffer from this disease. Please visit the hospital for further diagnosis and please take care of yourself.”

**Discussion and Results**

**Explanatory Analysis and unsupervised learning approach**

While exploring the dataset, we examined that the disease's variable was imbalanced with 148 different diseases from which further diseases were combined and reduced to 60 diseases, as shown in *Figure 5\_EDA* (Appendix). The analysis was conducted to investigate the intensity of diseases based upon other variables. From *Figure 4\_EDA* (Appendix), we examined that low and medium-intensity diseases are pretty close by, but the highest intensity diseases have abnormal BMI levels of people, reflecting the age being lower (higher weight). The distribution of the Age variable in the dataset is a left-skewed plot, with most data of ages 20 - 30. We also have data on teenagers and senior citizens. The healthy BMI level is between 19 to 21. We can examine an obesity level as BMI is greater than 25 for most of the data; it is somewhat normally distributed; *Figure 3\_EDA (Appendix)* top diseases to people older than 40 were psychotic disorder, kidney failure, respiratory infection, and malignant neoplasms. The malignant neoplasm diseases were higher in senior males as compared to females. We also examined what kind of similar symptoms some diseases had, one was fever was common in 40 different diseases from 148.

After reducing the number of diseases to 60, as discussed in the methods section, we further dig into the exploratory analysis of the clean data. We used multiple correspondence analysis techniques in our data in order to understand how categorical variables are related and further to analyze patterns of categorical variables. From *Figure 8\_MCA(Appendix)*, We observed that one disease has multiple symptoms so the figure is unclear, however, we can see some grouping of similar symptoms that predict some type of disease, but it is vague. Then implemented K Modes to further analyze the categorical variables. The KModes clusters the categorical variables, it uses the dissimilarities between the data points. The lesser the dissimilarities the more similar data points are.

We performed K Modes and MCA as a further dig deeper into the dataset to get useful insights. For KModes in *Figure 9\_Kmode(Appendix),*Based on the Elbow method we applied 3 clusters to variables age, weight and height. It was interesting to see the variables group together in 3 clusters. However, when we performed KModes on a symptoms variable, we didn’t get good results because we have a lot of data. The plot was not readable or visually appealing, thus we decided not to use it for further analysis.

In Multiple Correspondence Analysis in *Figure 8\_MCA(Appendix),* the relation between the categorical variables didn’t come out clear in plots. Similar symptoms were grouped together but didn’t point to any particular disease indicating which disease they belong to. Some symptoms like “choke”, “cardiovascular event” and “tonic seizures” are together that may indicate heart disease.

**Model building process and evaluation**

We used 4 different machine learning algorithms to predict the response variable “disease” based on the input. Initially, we created a model using an unbalanced dataset, only Xgboost and decision tree performed well on test data, and the rest other models did not perform well. We observed overfitting of training data. Then we created the SMOTE technique to balance the dataset by oversampling the minority class of the target variable ( “Disease”). All the models performed well on predicting disease. The results from Decision Tree were 0.96 accuracy on the testing set while 1.0 on the training set. From Random Forest, we got 0.92 on the testing set and 1.0 on the training set. Gradient Boosting on the training was 0.998 and got 0.88 accuracies on test data. The highest accuracy we got from XGboost was 0.966 and 1 from testing and training, respectively. *Figure 12 and Figure 13 (Appendix)*

We also started with Azure Cloud Auto ML to verify results and approaches. We got the highest accuracy from Auto ML from LightGBM Classifier of 0.85 and 75 on testing and training respectively. We used the hyperparameter tuning from Auto ML XGBoost on our manual XGboost model. The Auto ML did not outperform our manual model building as it did not cater to the imbalanced dataset. It made Count vectorization of symptoms variables. However, this was unnecessary as we had symptoms with one variable and whole sentences. If we analyze the Auto ML graph in terms of feature importance from *Figure 10\_Azure*, based on the final model, the Light Classifier was mostly influenced by the variables Height, Weight, Age, and Intensity of the diseases, and less importance was given to the symptoms themselves. As our dataset was multi-class classification and the model did not cater to imbalance, the main focus to analyze such a model was through the Precision and Recall graph *Figure 11\_Azure (Appendix)*. From the graph below in the appendix, we can examine that it had an F1 score of 81 % but the Precision and Recall graph was close to ideal only for the variables that had more data.  We used our domain knowledge and expertise to outperform our local model as compared to AutoML as we had domain knowledge of the dataset. We observed that implementing hyperparameter tuning on a decision tree model reduces the overall accuracy of the training set to 52% and test set to 51%, whereas increasing the performance of the Xgboost classifier model on both test and train data with an accuracy of  0.959 on test and 0.999 on training data. The overfitting got resolved after implementing parameter tuning in Xgboost.

|  |  |
| --- | --- |
| **Training set accuracy:** | **Testing set accuracy:** |
| **Imbalance dataset**  'DecisionTree': 1.0,  'GradientBoosting': 0.99,   'RandomForest': 1.0,   'XGboost': 0.96 | **Imbalance dataset**  'DecisionTree': 0.80,  'GradientBoosting':0.49  'RandomForest': 0.28,  'XGboost': 0.81 |
| **Balance dataset**  'GradientBoosting': 0.99,   'XGboost': 1.0,   'DecisionTree': 1.0,   'RandomForest': 1.0 | **Balance dataset**  'GradientBoosting': 0.87,   'XGboost': 0.96,   'DecisionTree': 0.95,   'RandomForest': 0.92 |
| **Hyperparameter tuning using balance dataset**  DecisionTree : 0.52  XGboost : 0.99 | **Hyperparameter tuning using balanced dataset**  DecisionTree : 0.50  XGboost : 0.95 |

*Figure 6 Analysis of testing and training accuracy from manual model building*

**Conclusion and Further Analysis**

The different techniques were implemented to predict disease based on age, weight, height, gender, the intensity of pain, the severity of symptoms and symptoms of an individual to diagnose the disease at an early stage. The XGboost classifier model with hyperparameter tuning performed well and was used for our front-end application. Web applications help users to predict disease settings in their comfort zone. Further help doctors to diagnose disease based on the medical history of the patient.

We acknowledge the fact that diseases cannot be detected purely with just one symptom but most of the datasets out for public use refer to just symptoms and disease data only. However, as only one symptom could not analyze the true diseases, having more symptoms cannot also decide person-specific diseases, this is due to the fact that each body/patient has different conditions in terms of age, BMI, and other health risk factors that cause specific diseases, they cannot be the same for each person. As mostly personal health records data is private and illegal to share publicly, we found it difficult to conclude our knowledge and expertise on a limited dataset.

For further analysis and research, we can apply our knowledge and expertise to a dataset that has a broader domain in terms of diseases, symptoms, and factors affecting those diseases on patients. As in real-world analysis and publications, one dataset is not always the key. We would like to focus on the data collecting part in order to combine different datasets that have demographic information along with the symptoms and diseases. This would also help us to generalize well on how different symptoms affect different body types.

**Appendix**

**Dataset**

Disease Prediction through Symptoms -  OHAS Dataset. Retrieved from

https://www.kaggle.com/usamag123/disease-prediction-through-symptoms

**Literature Review**

Article 1 [1] Efficient Automated Disease Diagnosis Using Machine Learning Models Naresh Kumar , 1 Nripendra Narayan Das,2 Deepali Gupta , 3 Kamali Gupta,3 and Jatin Bindra1. Retrieved from [https://www.hindawi.com/journals/jhe/2021/9983652/](https://arxiv.org/abs/2106.02813)

Article 2: [2] Rajora H, Punn NS, Sonbhadra SK, Agarwal S. Machine learning equipped web based disease prediction and recommender system. arXiv preprint arXiv:2106.02813. 2021 Jun 5. Retrieved from<https://arxiv.org/abs/2106.02813>

Article 3: [3] Keniya R, Khakharia A, Shah V, Gada V, Manjalkar R, Thaker T, Warang M, Mehendale N. Disease prediction from various symptoms using machine learning. Available at SSRN 3661426. 2020 Jul 27. Available at SSRN: <https://ssrn.com/abstract=3661426> or [http://dx.doi.org/10.2139/ssrn.3661426](https://dx.doi.org/10.2139/ssrn.3661426)

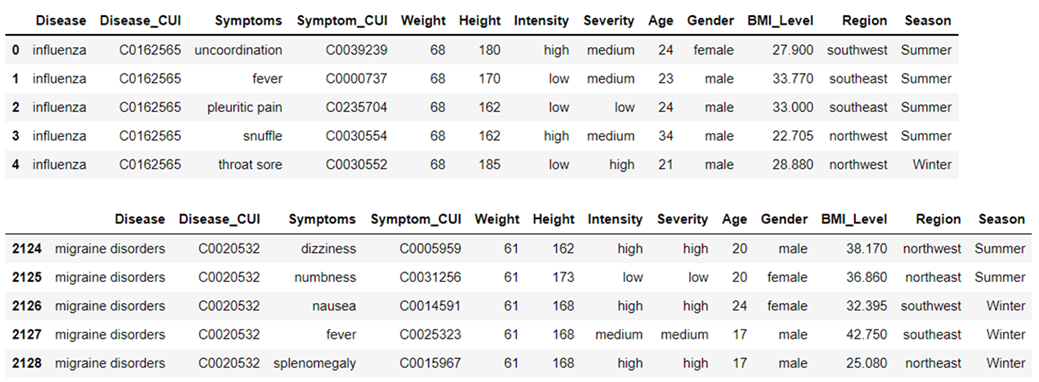
Article 4 : [4] Mathew RB, Varghese S, Joy SE, Alex SS. Chatbot for disease prediction and treatment recommendation using machine learning. In2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) 2019 Apr 23 (pp. 851-856). IEEE. Retried from https://ieeexplore.ieee.org/abstract/document/8862707

Article 5 : [5] Tamizharasi B, Livingston LJ, Rajkumar S. Building a medical chatbot using support vector machine learning algorithm. InJournal of Physics: Conference Series 2020 Dec 1 (Vol. 1716, No. 1, p. 012059). IOP Publishing. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1716/1/012059/meta

Article 6: [6] Doctor Chatbot: Heart Disease Prediction System 1Sherwin fernandes, 2Rutvij Gawas, 3Preston Alvares, 4Macklon Fernandes, 5Deepmala Kale and Shailendra Aswale. Retrieved from: http://www.iteejournal.org/v9no5oct20\_pdf12.pdf

Article 7 : [7] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE access. 2019 Jun 19;7:81542-54. Retrieved from https://ieeexplore.ieee.org/abstract/document/8740989

**Figure and Tables**

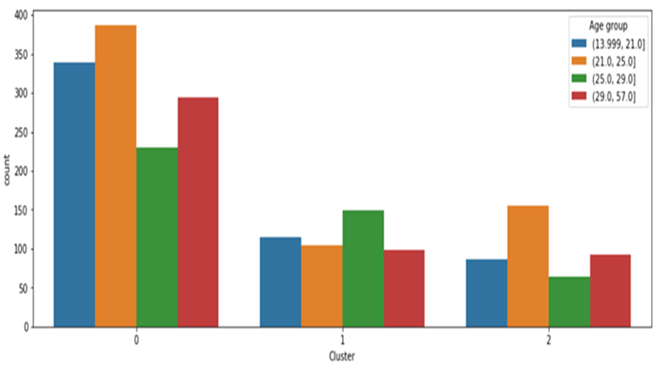


*Figure 1 Data overview*

Text

Description automatically generated with medium confidence

*Figure 9\_Kmode Data overview*

**

*Figure 9\_Kmode Output*

Chart, bar chart

Description automatically generated

*Figure 3\_EDA: Figure showing top common diseases in people older than 40*

Chart, box and whisker chart

Description automatically generated

*Figure 4\_EDA: Gender overview of symptoms intensity*

Chart, histogram

Description automatically generated

*Table 1\_Age\_distribution*

Chart, bar chart, histogram

Description automatically generated

*Figure 5\_EDA: Imbalance dataset plot after merging different diseases*

Chart, bar chart

Description automatically generated

*Figure 7\_EDA: Intensity count*

Chart

Description automatically generated

*Figure 10\_Azure: Azure LGBMClassifier Feature Importance*

Chart, line chart

Description automatically generated

*Figure 11\_Azure: Precision- Recall Curve*

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

*Figure 8\_MCA*

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated  *Figure 12: Accuracy of all classifiers using unbalanced dataset* | Chart, bar chart  Description automatically generated  *Figure 13: Accuracy of all classifier using balanced dataset* |