Symptom Based Disease Predictor and Diet Suggester

Under the Guidance of
Prof. V.K. Khatavkar

Dept. of Computer Engineering & IT
College of Engineering Pune (COEP)
May 2021

Project Members

Atharva Dixit (111703067) Computer Engineering
Gaurav Mandke (111703071) Computer Engineering
Abhijeet Chavan (141803001) Computer Engineering

Overview

1. Abstract
2. Problem Statement and Objective
3. Literature Review
4. Proposed Methodology/Solution
5. Experimental Setup
   - Hardware Requirements
   - Software Requirements
   - Setup
6. Results and Discussion
7. Conclusion
8. Future Scope
9. Publications
10. References

Abstract

- Biomedical and healthcare domains consist of a huge volume of unstructured data.
- The current disease prediction system doesn't ask any follow-up questions.
- It also fails to display a dietary remedy for the predicted disease.
- Our project necessitates using ensemble learning and/or feature selection on supervised learning algorithms.

Problem Statement and Objective

- **Problem**: Need of a system that can simplify the job of disease identification based on symptoms, with high accuracy for best results, while also providing a suitable diet plan to tackle the disease.
- **Solution:** The system is fed with datasets which are scraped from the web, so that the train-test split is carried out on pre-verified data. Enabling to ask co-occurring symptoms apart from those inputted.
- **Objective:** Developing a prediction system which will show the top possible diseases, and hence suggest the diet.
- Improving the prediction accuracy more by applying '**Ensemble Learning**' techniques and/or '**Feature Selection**' methods for data analysis.

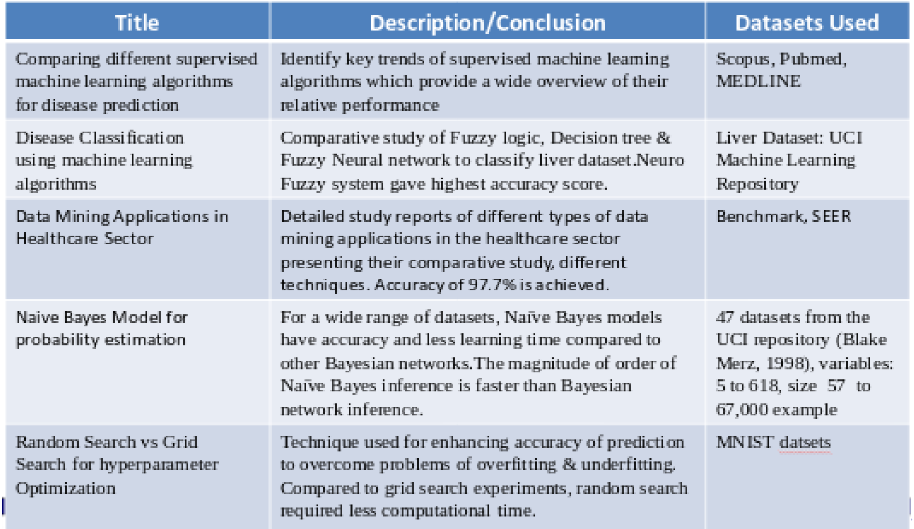
Literature Review

Based on the the segregation between research point of view and application point of view in the project, and the type of techniques used in analysing the phases can be divided into:

1. Finding out the most efficient and/or accurate machine learning model for predicting disease.
2. Improving accuracy of previously concluded methodology by ensemble learning and/or feature selection.
3. Suggesting Diet for predicted disease with respect to application point of view.

Literature Review

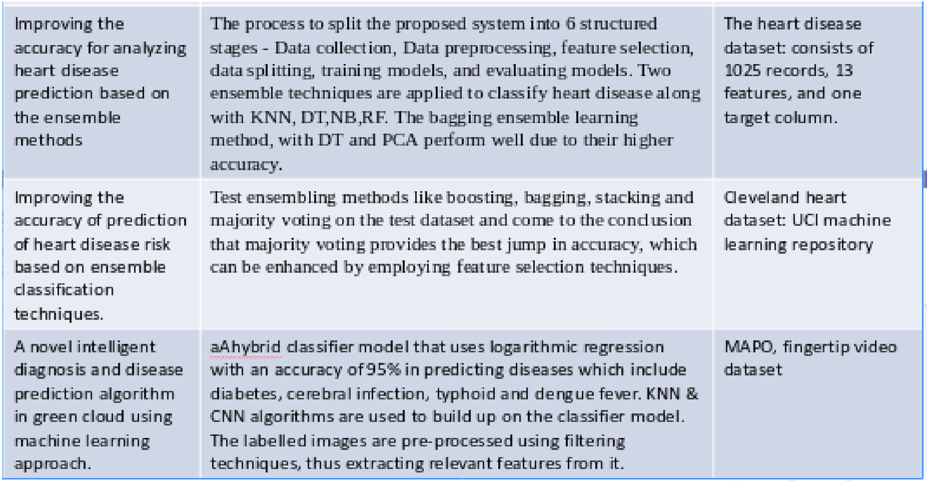
Finding out the most efficient and/or accurate machine learning model for predicting disease.



Title	Description/Conclusion	Datasets Used
Comparing different supervised machine learning algorithms for disease prediction	Identify key trends of supervised machine learning algorithms which provide a wide overview of their relative performance	Scopus, Pubmed, MEDLINE
Disease Classification using machine learning algorithms	Comparative study of Fuzzy logic, Decision tree & Fuzzy Neural network to classify liver dataset.Neuro Fuzzy system gave highest accuracy score.	Liver Dataset: UCI Machine Learning Repository
Data Mining Applications in Healthcare Sector	Detailed study reports of different types of data mining applications in the healthcare sector presenting their comparative study, different techniques. Accuracy of 97.7% is achieved.	Benchmark, SEER
Naive Bayes Model for probability estimation	For a wide range of datasets, Naïve Bayes models have accuracy and less learning time compared to other Bayesian networks.The magnitude of order of Naïve Bayes inference is faster than Bayesian network inference.	47 datasets from the UCI repository (Blake Merz, 1998), variables: 5 to 618, size 57 to 67,000 example
Random Search vs Grid Search for hyperparameter Optimization	Technique used for enhancing accuracy of prediction to overcome problems of overfitting & underfitting. Compared to grid search experiments, random search required less computational time.	MNIST datsets

Literature Review

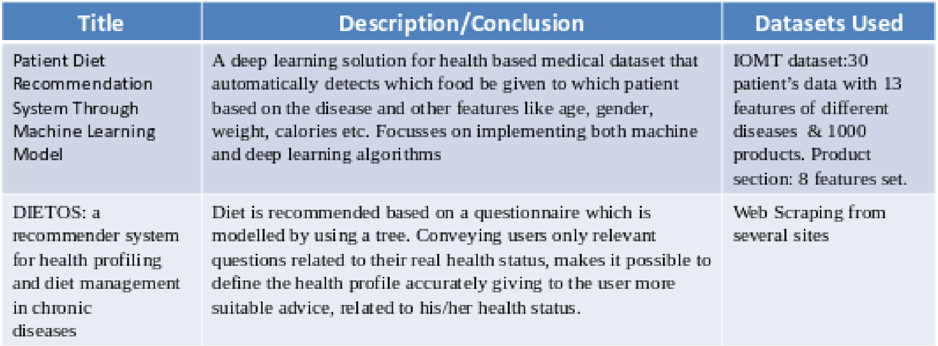
Improving accuracy of previously concluded methodology by ensemble learning and/or feature selection.



Improving the accuracy for analyzing heart disease prediction based on the ensemble methods	The process to split the proposed system into 6 structured stages - Data collection, Data preprocessing, feature selection, data splitting, training models, and evaluating models. Two ensemble techniques are applied to classify heart disease along with KNN, DT,NB,RF. The bagging ensemble learning method, with DT and PCA perform well due to their higher accuracy.	The heart disease dataset: consists of 1025 records, 13 features, and one target column.
Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.	Test ensembling methods like boosting, bagging, stacking and majority voting on the test dataset and come to the conclusion that majority voting provides the best jump in accuracy, which can be enhanced by employing feature selection techniques.	Cleveland heart dataset: UCI machine learning repository
A novel intelligent diagnosis and disease prediction algorithm in green cloud using machine learning approach.	aAhybrid classifier model that uses logarithmic regression with an accuracy of 95% in predicting diseases which include diabetes, cerebral infection, typhoid and dengue fever. KNN & CNN algorithms are used to build up on the classifier model. The labelled images are pre-processed using filtering techniques, thus extracting relevant features from it.	MAPO, fingertip video dataset

Literature Review

Suggesting Diet for predicted disease:application point of view.



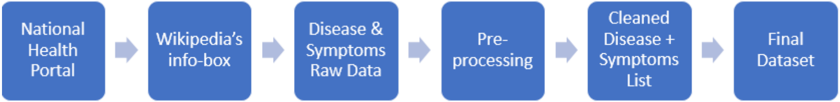
Title	Description/Conclusion	Datasets Used
Patient Diet Recommendation System Through Machine Learning Model	A deep learning solution for health based medical dataset that automatically detects which food be given to which patient based on the disease and other features like age, gender, weight, calories etc. Focusses on implementing both machine and deep learning algorithms	IOMT dataset:30 patient's data with 13 features of different diseases & 1000 products. Product section: 8 features set.
DIETOS: a recommender system for health profiling and diet management in chronic diseases	Diet is recommended based on a questionnaire which is modelled by using a tree. Conveying users only relevant questions related to their real health status, makes it possible to define the health profile accurately giving to the user more suitable advice, related to his/her health status.	Web Scraping from several sites

Datasets

1. Initial consideration: Kaggle datset having 132 symptoms as features and 4920 rows as target labels for diseases.
2. For a more accurate prediction and wide range of data, a more large and diverse dataset was formed by scraping data of size 8835 rows and 489 columns.
3. Severity Dataset from Kaggle for severity of disease/symptoms

Proposed Methodology/Solution

- Dataset is built using web scraping as shown below:



National Health Portal → Wikipedia's info-box → Disease & Symptoms Raw Data → Pre-processing → Cleaned Disease + Symptoms List → Final Dataset

Proposed Methodology/Solution

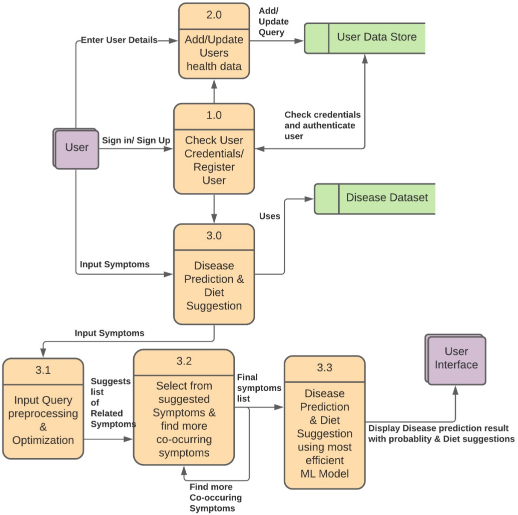

2.0
Add/Update Users health data
Add/ Update Query
User Data Store
Enter User Details
1.0
Check User Credentials/ Register User
Check credentials and authenticate user
User
Sign in/ Sign Up
Disease Dataset
Uses
3.0
Disease Prediction & Diet Suggestion
Input Symptoms
Input Symptoms
3.1
Input Query preprocessing & Optimization
Suggests list of Related Symptoms
3.2
Select from suggested Symptoms & find more co-ocurring symptoms
Final symptoms list
3.3
Disease Prediction & Diet Suggestion using most efficient ML Model
User Interface
Display Disease prediction result with probablity & Diet suggestions
Find more Co-occuring Symptoms

Figure: DFD Level 2

Proposed Methodology/Solution

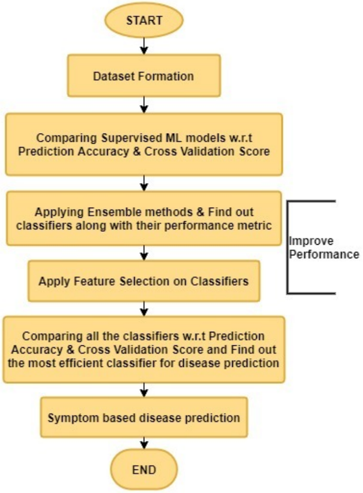

START
Dataset Formation
Comparing Supervised ML models w.r.t Prediction Accuracy & Cross Validation Score
Applying Ensemble methods & Find out classifiers along with their performance metric
Apply Feature Selection on Classifiers
Improve Performance
Comparing all the classifiers w.r.t Prediction Accuracy & Cross Validation Score and Find out the most efficient classifier for disease prediction
Symptom based disease prediction
END

Figure: Disease Prediction Methodology

Experimental Setup

1. Hardware Requirements
    - 4 GB RAM
    - 1 GHZ Quad Core Processor
    - 25 GB Hard Drive Space (for OS) + 2GB space for software dependencies
2. Software Requirements
    - Python 3.6 Above
    - Python Flask
    - Prominent Libraries
        1. NumPy, Pandas, Seaborn, Matplotlib
        2. Scikit Learn
        3. Beautiful Soup
        4. NLTK

Experimental Setup

1. Setup:
    1. Linux Machine
    2. Jupyter Notebook
    3. Linux Shell
    4. Any browser

Results and Discussion

- Comparison done between supervised machine learning algorithms along with RF Ensemble Model and ANN-MLP based on their accuracy.
- Cross validation scores were also compared and the model having the highest cross validation score used for predicting the disease.

Model	**Score(%)**
DT	83.60
LR	89.19
KNN	87.03
SVM	88.62
MNB	84.50
RF	87.13
MLP	86.77

Table: Model vs Score

Model	**Accuracy(%)**
DT	91.29
LR	90.72
KNN	91.29
SVM	90.05
MNB	83.94
RF	90.05
MLP	90.72

Table: Model vs Accuracy

Results and Discussion

- Applying ensemble learning techniques and feature selection methods we fulfil the aim of the project.
- Results to show a model more accurate and efficient than previous ones to serve purpose of the research using:

1. Ensemble Techniques
2. Feature Selection

Results and Discussion: Ensemble Techniques

- Bagging: increased accuracy as well as cross-validation score by upto 3-4 for most of the models

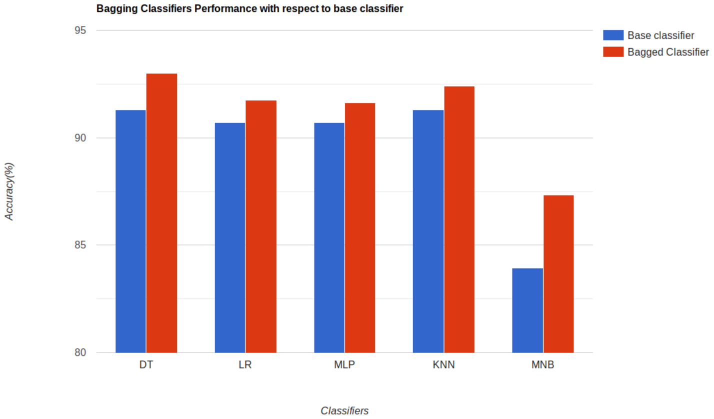


Figure: Bagging Classifier Performance w.r.t. base classifier

Results and Discussion: Ensemble Techniques

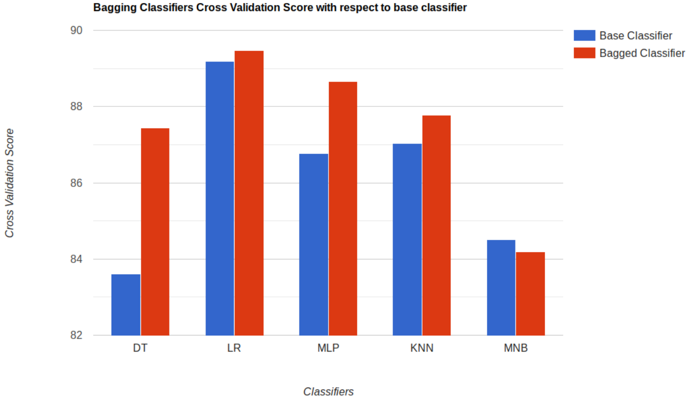


Figure: Bagging Classifier Cross Validation Score w.r.t. base classifier

- Boosting: only DT was used for experimentation. The Adaboost model increased accuracy as well as cross-validation score.

Results and Discussion: Ensemble Techniques

- XGB classifier:no increase in the accuracy.
- increasing order of models based on average accuracy (normal + cross validation): MNB, DT, MLP, RF, KNN, SVM, LR
- Majority Voting: incorporates several classifiers to increase their accuracy.We considered 3 strong models to form an ensemble with the remaining 4 weak.
  - Ensemble 1: LR + SVM + KNN + MNB
  - Ensemble 2: LR + SVM + KNN + MLP
  - Ensemble 3: LR + SVM + KNN + RF
  - Ensemble 4: LR + SVM + KNN + DT
- improves the accuracy of following weak classifiers by ensembling with strong classifiers: LR, SVM, KNN by a certain '%':
  - MNB : Accuracy by 7.92%, cross-validation score by 4.23%.
  - MLP : Accuracy by 0.23%, cross-validation score by 1.9%.
  - RF : Accuracy by 2.14%, cross-validation score by 1.45%.
  - DT : Accuracy by 0.68%, cross-validation score by 4.69%.

Results and Discussion: Ensemble Techniques

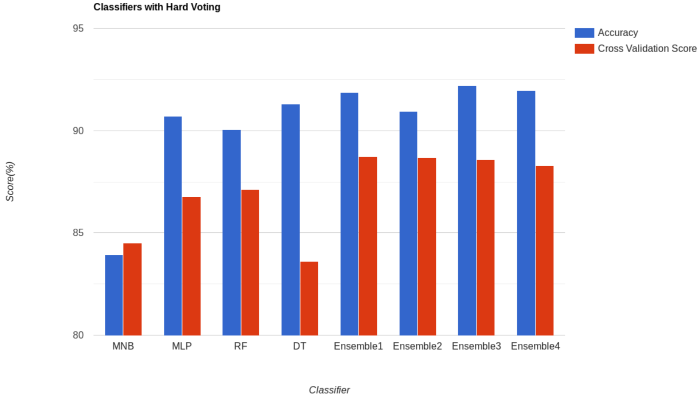


Figure: Classifier with Hard Voting

Results and Discussion: Ensemble Techniques

- Soft Voting:accuracy for weak classifiers increases more significantly, compared to hard voting. But the jump in the cross-validation score is smaller as compared to majority voting.
- Ensembling MNB, MLP, RF and DT with previous strong classifiers gave % :
  - MNB: accuracy by 8.59%, cross validation by 3.32%
  - MLP: accuracy by 2.27%, cross validation by 1.04%
  - RF: accuracy by 2.37%, cross validation by 0.97%
  - DT: accuracy by 1.02%, cross validation by 2.66%

Results and Discussion: Ensemble Techniques

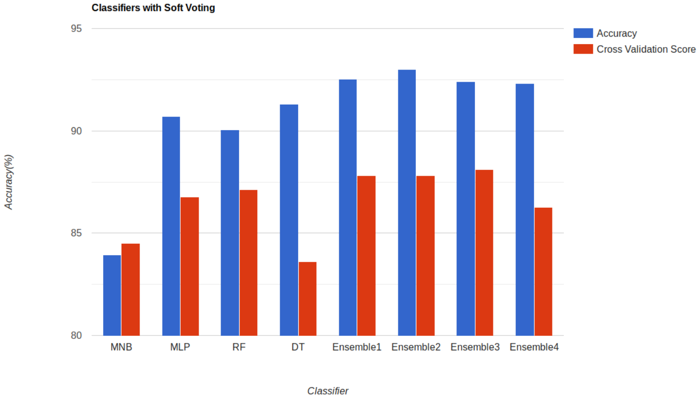


Figure: Classifier with Soft Voting

Results and Discussion: Ensemble Techniques

- Stacking: LR is used as a meta classifier. KNN, SVM, RF, MLP and DT combinations of these classifiers are used as base learners

Base Learner	Accuracy(%)	Score(%)
KNN+SVM+RF	89.82	84.32
KNN+SVM+RF+MLP+DT	90.27	87.09
KNN+SVM+MLP	90.27	86.29
KNN+SVM+RF+MLP+LR	90.38	86.89

Table: Stacking Accuracy and Cross Validation Score with LR meta classifier

Results and Discussion: Ensemble Techniques

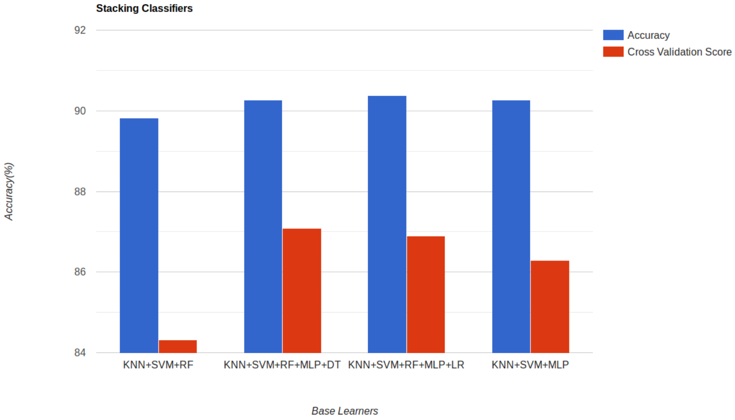


Figure: Stacking Classifier

Feature Selection

Techniques	Features Removed
Variance Threshold	5
Pearson Correlation	0
Chi-Squared Method	0
Mutual Information	0

Table: Different Feature Selection Techniques vs Number of unimportant Features removed.

- Combining Feature Selection Technique to use feature set with best performing ensemble classifier gave the desired results.
- The highest jump in cross-validation score: 0.18% for soft voting ensemble of LR+SVM+KNN+MLP
- the highest cross validation score: 89.50% for bagged-logistic regression model.

Bagged Model	Before FS(%)	After FS(%)
LR	89.47	89.50
DT	87.44	87.52
MLP	88.66	88.75
KNN	87.78	87.65
MNB	84.19	84.25

Table: Comparison of Cross-Validation Score(%) for Bagged Classifiers before and after Feature Selection

Hard Voting Ensemble	Before FS(%)	After FS(%)
LR+SVM+KNN+MNB	88.73	88.73
LR+SVM+KNN+MLP	88.67	88.73
LR+SVM+KNN+RF	88.58	88.64
LR+SVM+KNN+DT	88.29	88.25

Table: Comparison of Cross-Validation Score(%) for Hard Voting Ensembles before and after Feature Selection

Soft Voting Ensemble	**Before FS(%)**	**After FS(%)**
LR+SVM+KNN+MNB	87.82	87.88
LR+SVM+KNN+MLP	87.81	87.98
LR+SVM+KNN+RF	88.10	88.15
LR+SVM+KNN+DT	86.26	86.23

Table: Comparison of Cross-Validation Score(%) for Soft Voting Ensembles before and after Feature Selection

Conclusion

- A non-skewed/non-sparse dataset plays a critical role in training the model.
- Wide domain can be ensured if a symptom can be associated to several diseases
- Majority Voting (Hard Voting) works well with the weak classifier set
- Logistic Regression and Bagging show the highest cross-validation score

Future Scope

1. Ability to accept factors like family history, blood test reports, age, gender, region-wise traits to link disease onset in case of genetic or hereditary diseases.
2. Using Medline dataset for non-numerical traits, further increasing system accuracy.
3. Generating different feature sets for the same using Feature Selection.
4. These sets can be useful in determining the most accurate predictive model.

Publications

Paper Title: **Enhancing Accuracy of Symptom-Based Disease Prediction using Ensemble Techniques and Feature Selection**

Submitted to:

3rd International Conference on Machine Intelligence and Signal Processing (MISP) 2021, NIT Arunachal Pradesh

References

- Shahadat Uddin ,Arif Khan: BMC Medical Informatics and Decision Making research paper “Comparing different supervised machine learning algorithms for disease prediction” BMC Med Inform Decis Mak 19, 281 (2019).

- International Research Journal of Engineering and Technology (IRJET), A Smart Health Prediction Using Data Mining, e-ISSN: 2395-0056

- S.SHARMILA, International Journal of Advanced Networking Applications (IJANA), Vol: 08, Issue: 05, 2017, “Disease Classification using Machine Learning Algorithm”

- K. Vembandasamy, IJISET - International Journal of Innovative Science, Engineering Technology, Vol. 2 Issue 9, September 2015, “Heart Diseases Detection Using Naive Bayes Algorithm”

- Durai Raj V Ranjani Data Mining Applications in Healthcare Sector INTERNATIONAL JOURNAL OF TECHNOLOGY RESEARCH VOLUME 2, ISSUE 10, OCTOBER 2013

- Archana Singh , Rakesh kumar: IEEE a research paper ”Heart Disease Prediction Using Machine Learning Algorithms” publish in 2020 International Conference on Electrical and Electronics Engineering (ICE3)

References

G Agapito, B Calabrese, PH Guzzi, M Cannataro, M Simeoni, I Car´e, T Lamprinoudi, G Fuiano, and A Pujia.2016. DIETOS: A recommender system for adaptive diet monitoring and personalized food suggestion. In Wireless and Mobile Computing, Networking and Communications (WiMob), 2016 IEEE 12th International Conference on. IEEE, 1–8

Perumal Venkatesan,Leoni Sharmila:”International Journal of Pure and Applied Mathematics” on Project”Machine Learning Approaches for Biological Data mining”research paper ”Disease Classification Using Machine Learning Algorithms-A Comparative Study”

https://www.hindawi.com/journals/complexity/2021/6663455/

https://link.springer.com/article/10.1007/s11227-020-03481-x

https://link.springer.com/chapter/10.1007/978-3-319-13728-5

http://www.ijecs.in/index.php/ijecs/article/view/1855

https://www.researchgate.net/publication/344438361

Questions and Discussions