Name and Surname: Mariam Kili Bechir

Chronic Diseases detection Using Machine Learning Techniques

1. Introduction

Chronic diseases prediction is a vital subject because it addresses a critical need in healthcare which is a major concern for humanity. Detecting chronic diseases early is crucial, and that's where advanced automation systems such as AI and Machine learning come into play [1]**.** The goal of this research is first of all finding a way to automatically predict chronic diseases based on patience's records without taking more time and by making doctor's tasks easier. It's also an important subject because this one combine healthcare domain with artificial intelligence. It's a good idea to modernize healthcare by using technology. Because nowadays with the growth of the domain of data and AI, the goal is to transform traditional systems into automated and more improved systems.

By harnessing the power of AI and Machine Learning, we aspire to create a system that not only enhances accuracy but also expedites the diagnostic process. The implications are profound, potentially leading to more effective treatments and improved patient outcomes.

According to existing research's, building chronic disease prediction using Machine Learning is essential because this subject not only include Machine Learning domain, but also healthcare field and Machine Learning project building pipeline. This project allows us to identify, preprocess and build successful models using some methodologies. By critically assessing what has been done, we pave the way for innovation, ensuring that our research is not only novel but also informed by the collective wisdom of the scientific community.

2. Organization of literature review

- [2] highlight the investigation of chronic diseases using data mining techniques. In this research published in 2015, it's mentioned that the correlation of various chronic diseases has become a necessity and this can be achieved by using data mining techniques, which help to derive knowledge about the effects of a particular chronic disease on the other chronic diseases.
- In their conference published in 2016, In their conference published in 2016, (Astya et al., n.d.) discuss about methods used to predict diagnostic codes for chronic diseases using machine learning techniques. They focused on 11 types of different chronic diseases such as kidney disease, osteoporosis, arthritis etc and used Machine learning techniques for each disease's diagnosis.In their conference published in 2016, (Astya et al., n.d.) discuss about methods used to predict diagnostic codes for chronic diseases using machine learning techniques. They focused on 11 types of different chronic diseases such as kidney disease, osteoporosis, arthritis etc and used Machine learning techniques for each disease's diagnosis.
- [4] prove how Machine-learning can improve the accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others.
- [5] In this paper published in 2018, (IEEE Circuits and Systems Society. India Chapter and Institute of Electrical and Electronics Engineers n.d.) five chronicle datasets are taken and the machine learning algorithms such as decision tree, random forest, and the support

vector machine are applied and the predicted whether the patient is suffering from a chronicle disease such as heart disease, liver disease or diabetes. A result was obtained by comparing all algorithms performance on all dataset the random forest predicts with high accuracy.

- [6] aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.
- [7] use Machine learning predictive models to predict chronic diseases. They precise that among the methods considered, support vector machines (SVM), logistic regression (LR), clustering were the most commonly used. These models are highly applicable in classification, and diagnosis of CD and are expected to become more important in medical practice in the near future.
- Logistic regression was as good as machine learning for predicting major chronic diseases [8] to evaluate the performance of machine learning (ML) algorithms and to compare them with logistic regression for the prediction of risk of cardiovascular diseases (CVDs), chronic kidney disease (CKD), diabetes (DM), and hypertension (HTN) .
- [9] demonstrates the noncrisp Rough K-means (RKM) clustering for figuring out the ambiguity in chronic disease dataset to improve the performance of the system. The experimental results demonstrate that the proposed system is successfully employed for the diagnosis of chronic diseases. The proposed model achieved the best results with naive Bayes with RKM for the classification of diabetic disease (80.55%), whereas SVM with RKM for the classification of kidney disease achieved 100% and SVM with RKM for the classification of cancer disease achieved 97.53 with respect to accuracy metric.
- In [10] study, a comparison between deep learning model and twelve machine learning and ensemble learning methods based on relatively small data including 183 healthy individuals and 401 early Parkinson Disease patients shows the high detection performance of the designed model, with 96.45% of accuracy.
- [11] combines Deep Learning and Machine Learning techniques to predict heart disease. Using deep learning approach, 94.2% accuracy was obtained.
- In this study [12] Naïve Bayes and KNN algorithms was used. The accuracy of heart disease prediction using naive Bayes obtained was 94.5% which is greater than accuracy of KNN. They then compared naive Bayes with KNN and figure out that KNN requires more memory and time. Additionally, a risk prediction system using the CNN algorithm was developed to assess the risk of heart disease.
- [13] proposed a study that concerns the cardiac disease diagnosis. After applying to preprocess and feature engineering techniques, machine learning approaches like random forest, decision trees, gradient boosted trees, linear support vector classifier, logistic regression, one-vs-rest, and multilayer perceptron are used to perform binary and multiclassification on the dataset.
- In the study [14] published in 2022, a Deep learning model for multi-classification of infectious diseases was build based on unstructured electronic medical records in order to assist in clinical infectious-disease decision-making. The accuracy of MIDDM (Multi-classification of Infectious diseases Model) achieved is 99.44%, which is significantly higher than that of XGBoost (96.19%), Decision tree (90.13%), Bayesian method (85.19%), and logistic regression (91.26%).
- In this study, [15] various machine learning algorithms have been used in the training process to predict diseases belonging to different branches of medicine, such as diabetes,

bronchial asthma, and covid. It is also the first study to achieve an accuracy score of 99.33% with a dataset that involves a greater number of diseases.

From this literature review, we can conclude that researches based on the application of Machine Learning techniques in the healthcare domain continue to grow from each year to other. The analyses of these studies from 2015 to 2023 show us that we have a continuous grow in the applied Machine Learning and artificial intelligence techniques in the healthcare domain, especially in diseases diagnosis such as chronic diseases. This leads to a conclusion where we can say Machine Learning techniques including Deep Learning have a huge contribution in the development of healthcare.

3. Summary and Synthesis

In [3] , [5], [7], [8] , [13], [15] studies, Machine Learning classification algorithms such as Logistic Regression, Decision Tree, SVM and RandomForest algorithms was applied. In most of these studies, the most used technique is Support Vector Machine algorithm which is a good technique that can perform well in classification problems. These studies have all shown successful result on predicting models that will detect chronic diseases such as kidney disease, heart disease, covid, arthritis, diabetes, osteoporosis etc. The successful outcomes reported in these studies underscore the viability of Machine Learning algorithms for the detection and prediction of diverse chronic diseases.

The studies [11] and [14]] advanced the exploration by integrating both Machine Learning and Deep Learning techniques. Employing algorithms such as Logistic Regression, Decision Tree, Bayesian methods, and introducing Deep Learning techniques like auto-encoder deep learning, these studies presented a hybrid approach. [11] employed a sequential model featuring a fully connected dense layer with flatten and dropout layers to prevent overfitting. Intriguingly, the results obtained from Deep Learning techniques surpassed the performance of traditional Machine Learning algorithms. This superiority can be attributed to the inherent power of Neural Networks, highlighting their capacity to outperform conventional learning algorithms.

4. Conclusion

The literature review reveals a prevalent use of Machine Learning algorithms for chronic disease prediction. Among various algorithms SVM and Logistic Regression, stands out for its effectiveness in classification tasks related to chronic diseases. Hybrid approaches, integrating both Machine Learning and Deep Learning, show promise in enhancing prediction accuracy. The integration of Deep Learning techniques shows us the robust capabilities of Neural Networks. As chronic diseases pose significant global health challenges, accurate prediction models contribute to early detection and intervention, ultimately improving patient outcomes.

II- Preparing the Data Research

1. Introduction:

Our research project aiming to predict chronic diseases through machine learning techniques. For this, we need to collect data, preprocess them by applying Exploratory Data Analysis and then choosing appropriate machine learning models and training them on respective diabetes, heart disease, and Parkinson's disease datasets.

2. Organization

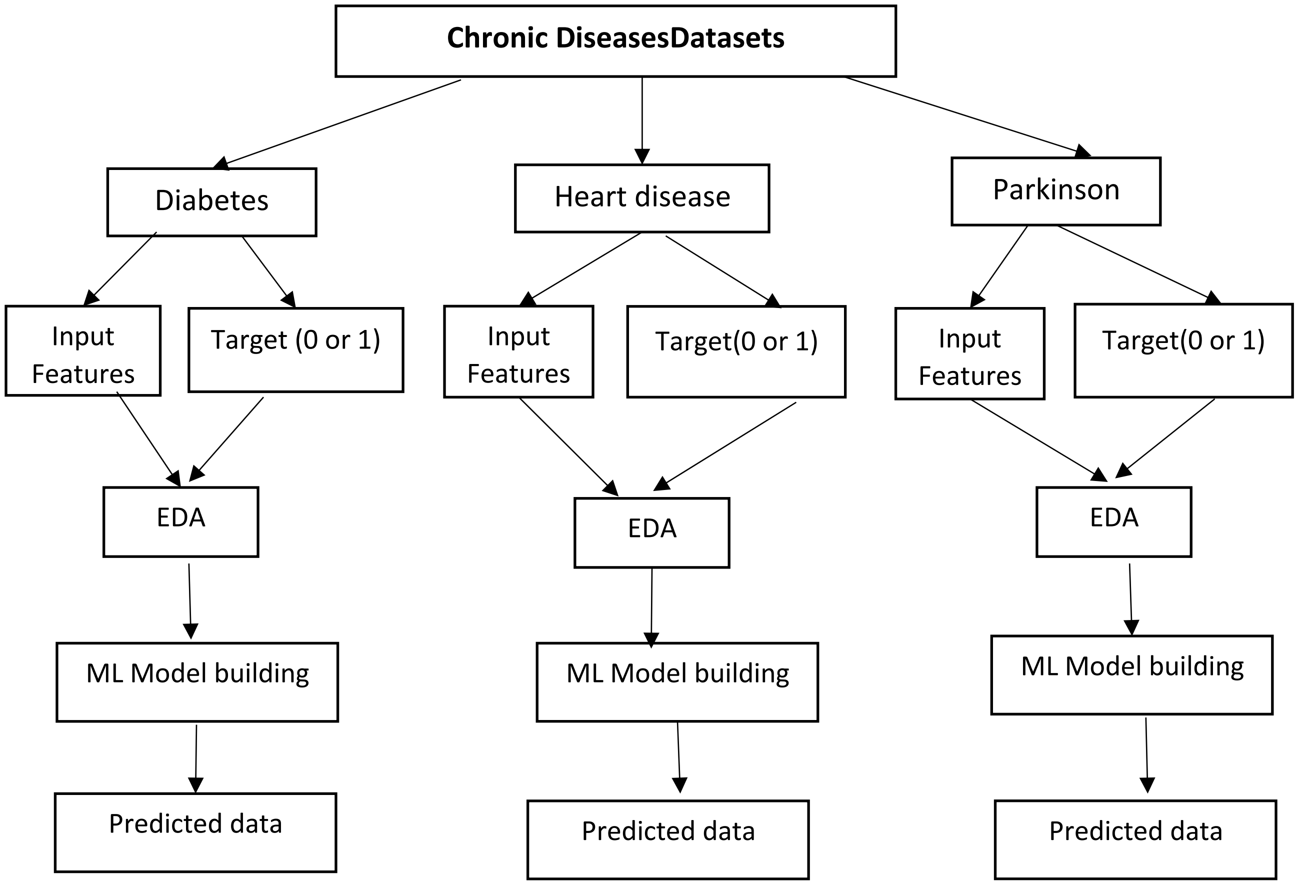

Figure 1data organization

3. Data Description

The data used for this study are collected from Kaggle which is an open-source platform for datasets. It’s a collection of three different datasets. Each dataset represents 1 type of chronic diseases. Totally we have 3 different chronic diseases which are: diabetes, heart disease and Parkinson diseases. Each dataset size is different from the other. The total of diabetes dataset is 768 entries and 8 independent features which are: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and the Outcome feature which is the target that will going to be trained as y_train and predicted later.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...

Figure 2-Sample of diabetes dataset

For heart disease, the total entries are 303 rows and we have 14 attributes in which one attribute is the target feature that need to be predicted.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3-Sample of heart disease dataset

About Parkinson disease, 195 entries are registred with 24 different features in which 23 are dependent and 1 feature represent the target independent feature.

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	...
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374	...
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	...
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233	...
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	...
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	...
...	...	...	...	...	...	...	...	...	...	...	...

Figure 4- Sample of Parkinson disease dataset

4. Data Analysis and Insights

In this study, each of the 3 datasets has a binary input which means all our targets have 0 and 1 values. To analyse these data, extract insights and make decision about these data, Exploratory Data Analysis(EDA) techniques can be used. EDA includes handling missing values if its exist, detecting outliers, analysing statistically each datasets, visualize data and make insights etc.

Diabetes diseases dataset contains totally 768 rows and 9 columns which represent the attribut es. The attributes that contains 0 as minimum value and 1 as maximum value is the target or t he output variable of our data.

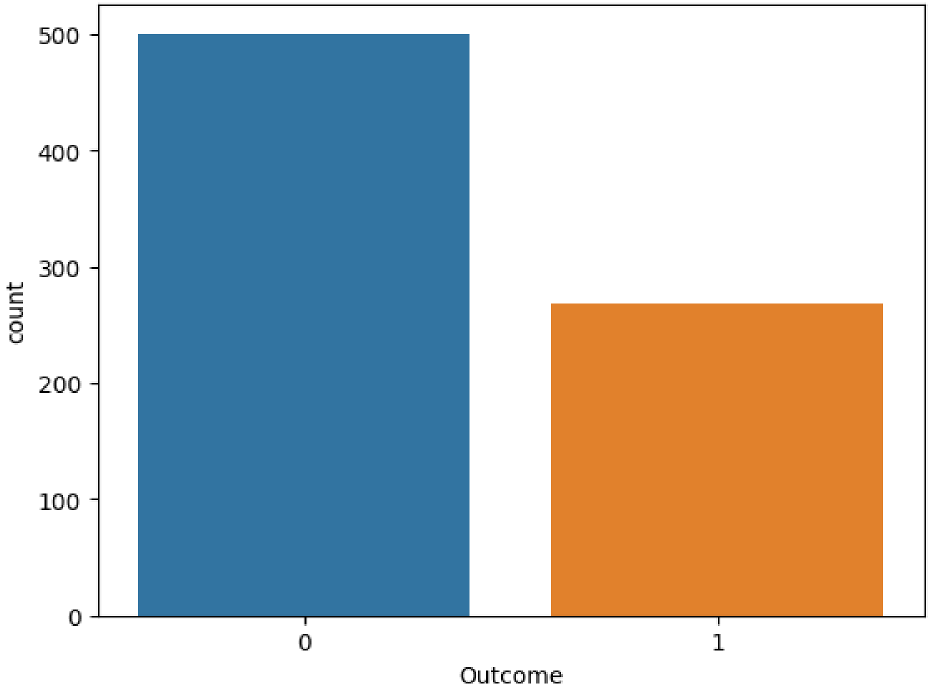

Figure 5- The output label insight of Diabetes disease

According to this graph, the person who are not diabetic are more than the person who have diabetes. Around 500 are non-diabetics and 268 persons are affected by diabetes.

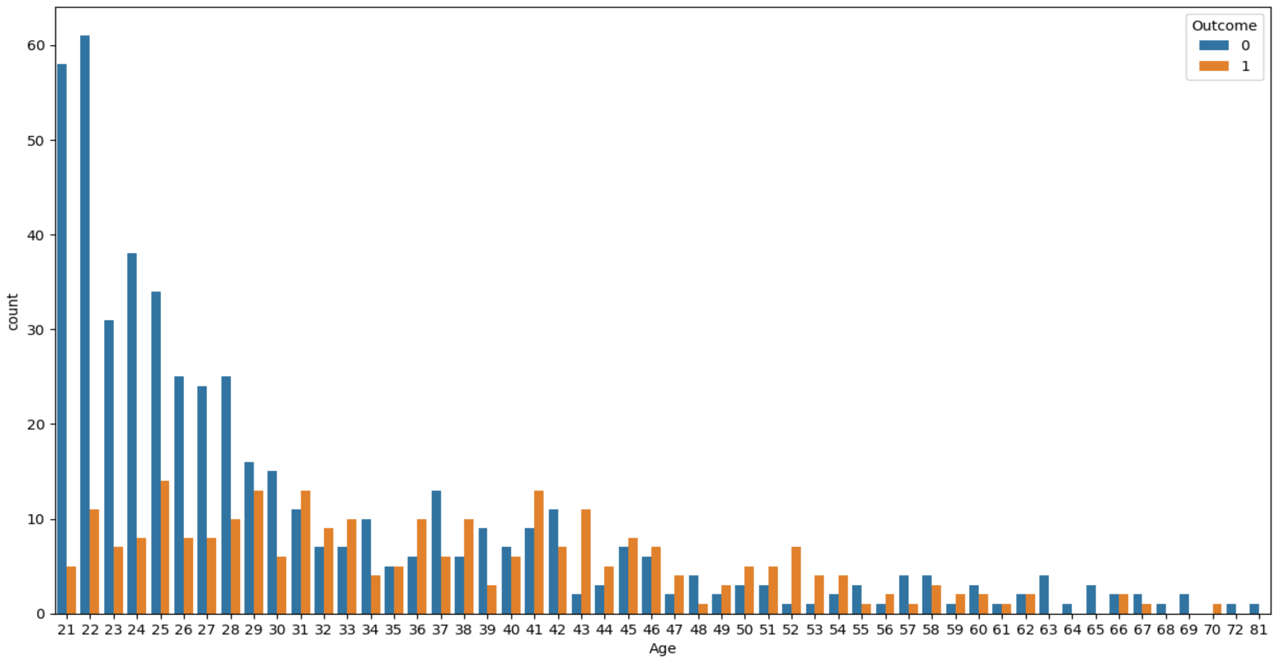

Figure 6- relationship between Ages and diabetes

This graph shows us that diabetes diseases is more shown in people who are more than 30 years old. People between 21 and 30 are less affected. That means that the more the age increase, the more people have risk to be more diabetics.

In heart diseases dataset, the total shape is 303 rows × 14 columns. The output label is a set of 0 and 1 values.

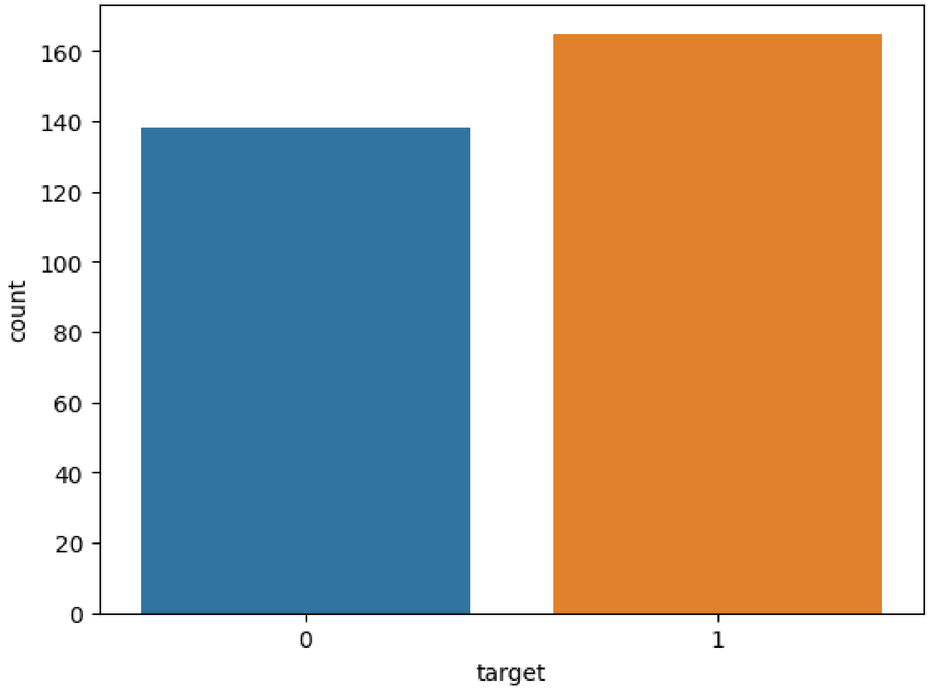

Figure 7- The output label insight of heart disease

The person affected by heart diseases are more than the persons non affected.

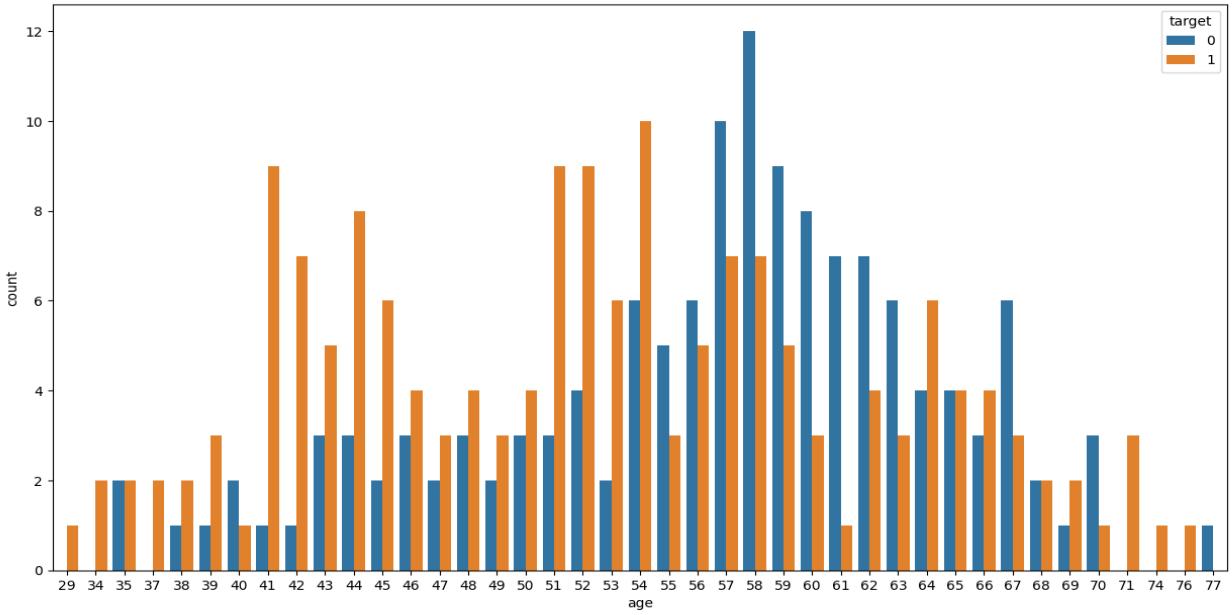

Figure 8relationship between Ages and heart disease

The insight shows that people between 40 and 60 are more likely to be affected by the heart disease and also people who are greater than 65 years.

Parkinson disease is also one of chronical disease that need to be analysed. In our study, the total entries are 195 and the the number of features is 24, these features are more medical features.

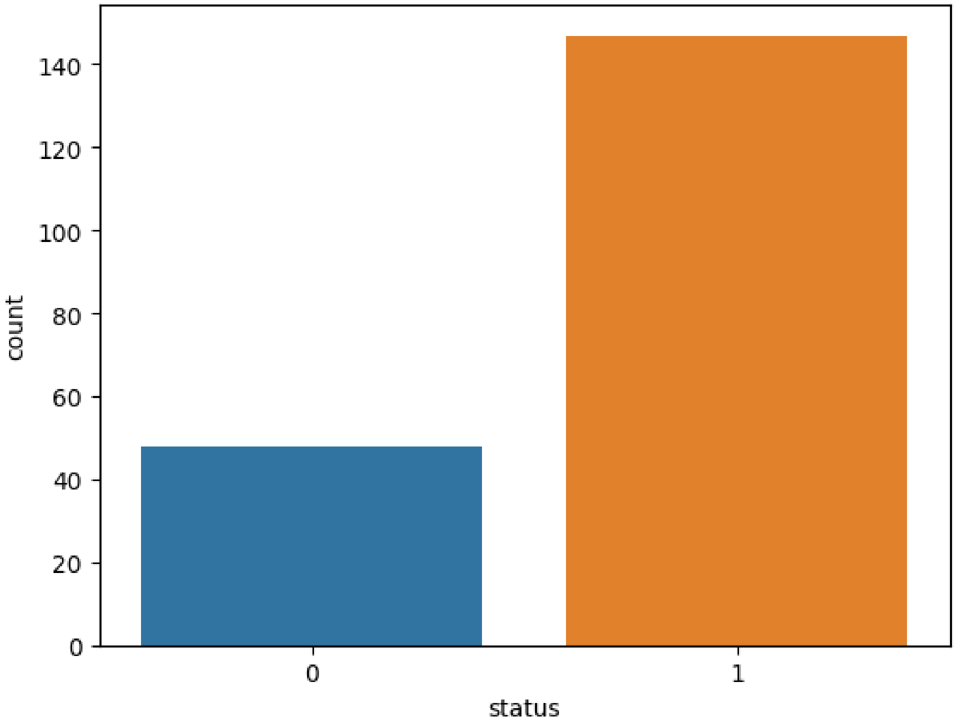

Figure 9- The output label insight of Parkinson disease

This is the countplot of the output label. 1 represents the person affected by Parkinson disease and 0 represents the persons non affected. The number of person affected by the disease is very high according to the result.

5. Conclusion

This research not only illuminates the complexity of chronic diseases but also highlight the importance of early detection. By using EDA techniques, we can extract more insights from the data. This is a good approach for future studies to use machine learning models to predict and mitigate these health challenges. The insights gained from our work contribute significantly to the existing body of knowledge, helping future researchers find even better solutions for healthcare.

Through meticulous data analysis, we move one step closer to empowering healthcare professionals with tools for proactive disease management and prevention.

6. Proper Citations

To predict diabetes diseases, [16] use a dataset that contains 800 records and 10 attributes. The dataset is then preprocessed and prepared for the model building. In [11] The dataset consists of 14 main attributes used for performing the analysis. The number of attributes used in this study are the same with our number of attributes in heart datasets. [10] use data including 183 healthy individuals and 401 early Parkinson Disease patients, use preprocessing techniques to analyze and preprocess the data and then build a deep learning model.

III- Preparing Technology Review

1. Technology Overview

Since the targets of our datasets are binary values (0 and 1), we should apply supervised Machine Learning techniques in order to build a classification model for prediction for each of our datasets. Binary classification algorithms such as Logistic Regression and also compare it with others Machine Learning classification algorithms such as SVM, Decision Tree, K-Nearest Neighbor Classifiers.

Logistic Regression is a supervised Machine Learning algorithms mostly used for binary classification problems to build a predictive model. If the output data that will going to be predicted is in form of 0 and 1, this algorithm can be used. There are others Machine Learning classifications such as Support Vector Machine, K-Nearest Neighbors and Decision Tree.

SVM is a powerful supervised algorithm that works best on smaller datasets but also on complex ones. Support Vector Machine, abbreviated as SVM is an algorithm that try to find a hyperplane that best separates the two classes. It can be used for both regression and classification tasks, but generally, they work best in classification problems. It's one of the most popular algorithm used to build a model for disease prediction when we have small or complex dataset with continuous values.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. This algorithm is Simple to understand and to interpret and it can also handle both categorical and numerical variables. With Decision Trees, it's Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.

Additionaly, we have also ensemble techniques such as RandomForest and GradientBoosting algorithms such as XGBOOST. Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and merges their predictions to improve the overall accuracy and control overfitting. It's a good algorithm to solve also classification problems and it's widely used in various domains due to its flexibility, robustness, and high performance. XGBoost (eXtreme Gradient Boosting) is an efficient and scalable machine learning algorithm that belongs to the family of gradient boosting frameworks. This algorithm is known for its high performance and effectiveness in various machine learning tasks, particularly in structured dataset.

These Machine Learning Techniques are necessary for our project for model building. After this the step of model building, this project needs to be implemented in a graphical user interface. So, a web application using Streamlit need to be used to build the entire system. Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science.

6. Relevance of Machine Learning Techniques to our Project

Predicting diseases is recognized as a classification problem. In our study, the outputs are 0 and 1. 0 represents the person not affected by the disease and 1 is the result of the affected disease. It's the case for all of our three different datasets. In this case, LogisticRegression seems to be the first choice to build our models. But we need also to use others classifications algorithms such as SVM or Decision Tree or also ensemble techniques to the Logistic Regression algorithm to chose the best accuracy. The models that perform well are the model that have the highest accuracy.

In addition to these model building techniques, Hyperparameter Tuning technique such as GridSearch will be used to optimize our models and chose the best ones.

To analyze the performance of our models, there are some metrics such as confusion matrix, accuracy score and classification reports that can help us to evaluate our models.

After building, evaluating and the different models, an interface will be used to build to build a Multiple disease prediction system. For this, we will use Streamlit to build the web application and then deploy the application built.

7. Comparison and Evaluation

When considering machine learning models for predicting chronic diseases, it's essential to assess various models based on their strengths, weaknesses, and interpretability. Here's an analysis for some commonly used models:

- Logistic Regression is simple and computationally efficient that provides probabilities for outcomes. One of the strengths of this algorithms is that it's robust to overfitting, especially with a small number of features. However, Logistic Regression assumes a linear relationship between features and the log-odds of the outcome and it may not perform well with highly non-linear data.
  Logistic Regression is highly interpretable as it directly models the probability of the output.
- Decision Trees: This algorithm is intuitive and easy to understand, making the model interpretable. It Can handle both numerical and categorical data and also performs automatic feature selection. However, Decision Tree can Prone to overfitting, especially on complex datasets and it's sensitive to small variations in the data.
- Support Vector Machines (SVM) is effective in high-dimensional spaces and versatile due to different kernel functions. It's also resistant to overfitting. But one of the weaknesses is that the computationally is hight, especially with large datasets and the model parameters (e.g., the choice of the kernel) can be challenging to interpret.
- Random Forest**:** The strength of RandomForest is that this algorithm is Robust and less prone to overfitting compared to individual decision trees. It can handle a large number of features and provides feature importance scores. The Weaknesses are Lack of interpretability compared to decision trees and the computation is sometimes more expensive.

  RandomForest is less interpretable compared to decision trees, but feature importance can be assessed.

- Neural Networks have the ability to capture complex relationships in data. It has an excellent performance on large and diverse datasets. However, Neural Networks require large amounts of data for training and prone to overfitting, especially on small datasets most of the time.

Considering the need for interpretability in predicting chronic diseases, a combination of Logistic Regression, SVM and Random Forests could be a suitable choice, providing a balance between accuracy and model transparency. Logistic Regression could be a strong candidate, especially when the goal is to understand the influence of individual features on the likelihood of a specific disease. However, the final choice should also consider the nature of the data, the specific requirements of the healthcare project, and the desired balance between model simplicity and predictive accuracy.

5. Use Cases and Example and citations

In this study [8] Logistic Regression was a good choice for building the machine learning model for predicting chronic diseases. Logistic regression reached the highest area under the receiver operating characteristic curve for chronic kidney diseases (0.905 [0.88, 0.93]) and Diabetes model (0.768 [0.73, 0.81]) predictions. SVM, Decision Tree and KNN are also a good types of Classification algorithms and can perform well depending on the type of datasets and the hyperparameters Tuning. In this study [15], Various machine learning algorithms have been used in the training process to predict diseases such as diabetes bronchial asthma, and covid using Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Multilayer Perceptron, Decision Tree, Random Forest algorithms and XGBoost. They achieve an accuracy score of 99.33%. This paper [17] prove that multi algorithms such as Decision Tree, Naïve Bayes, Logistic Regression, Random Forest, SVM was used for both structured and unshaped data. The Decision Tree algorithms gave the highest accuracy score, 93.24%.

With ensemble Techniques, we can also get a good accuracy for disease classification problems. Ensemble techniques such as RandomForest and Xgboost use multiple learning algorithms to obtain better predictive performance. In this study [18], an improved ensemble learning approach was developed for the prediction of heart disease risk on two different datasets. The first dataset achieved an accuracy of 83% for XGBoost and 84% for RandomForest. The second dataset gave an accuracy of 81% for XGBoost and 84% for RandomForest.

Deep Neural Network (DNN) and Artificial Neural Networks techniques are used by [19] to predicts the presence or absence of heart disease. The model obtained shows better results on both the testing and training data. [20] highlight the application of deep learning techniques in diseases prediction. They introduce several deep learning algorithms: Artificial Neural Network (NN), FM-Deep Learning, Convolutional NN and Recurrent NN.

By reviewing previous studies, we figure out that the methods used for building models to predict diseases are mostly classifications Machine Learning techniques, ensemble techniques and also deep learning techniques. [20] affirm that the future medical technology should be combined with Digital Twins to realize real intelligent medical treatment, pay more attention to personalized medical treatment, integrate with precision medical treatment, and serve individuals more conveniently.

[1] S. A. Bini, ‘Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?’, *J Arthroplasty*, vol. 33, no. 8, pp. 2358–2361, Aug. 2018, doi: 10.1016/J.ARTH.2018.02.067.

[2] Panjab University. University Institute of Engineering and Technology and Institute of Electrical and Electronics Engineers, *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS) : (December 21-22, 2015), University Institute of Engineering and Technology, Panjab University, Chandigarh, 160014, India.*

[3] P. N. Astya, Galgotias University. School of Computing Science and Engineering, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section. SP/C Joint Chapter, and Institute of Electrical and Electronics Engineers, *Proceeding, International Conference on Computing, Communication and Automation (ICCCA 2016) : 29-30 April, 2016.*

[4] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, ‘Can Machine-learning improve cardiovascular risk prediction using routine clinical data?’, *PLoS One*, vol. 12, no. 4, Apr. 2017, doi: 10.1371/JOURNAL.PONE.0174944.

[5] IEEE Circuits and Systems Society. India Chapter and Institute of Electrical and Electronics Engineers, *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET).*

[6] X. Liu *et al.*, ‘A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis’, *Lancet Digit Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019, doi: 10.1016/S2589-7500(19)30123-2.

[7] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, ‘Applications of machine learning predictive models in the chronic disease diagnosis’, *J Pers Med*, vol. 10, no. 2, Jun. 2020, doi: 10.3390/jpm10020021.

[8] S. Nusinovici *et al.*, ‘Logistic regression was as good as machine learning for predicting major chronic diseases’, *J Clin Epidemiol*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/J.JCLINEPI.2020.03.002.

[9] T. H. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, ‘Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms’, *J Healthc Eng*, vol. 2020, 2020, doi: 10.1155/2020/4984967.

[10] W. Wang, J. Lee, F. Harrou, and Y. Sun, ‘Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning’, *IEEE Access*, vol. 8, pp. 147635–147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[11] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, ‘Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning’, *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/8387680.

[12] S. Vilas and A. M. S. Scholar, ‘Diseases Prediction Model using Machine Learning Technique’, doi: 10.32628/IJSRST.

[13] D. Hamid, S. S. Ullah, J. Iqbal, S. Hussain, C. A. U. Hassan, and F. Umar, ‘A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream’, *J Sens*, vol. 2022, 2022, doi: 10.1155/2022/8400622.

[14] M. Wang, Z. Wei, M. Jia, L. Chen, and H. Ji, ‘Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records’, *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-01776-y.

[15] M. ÇOLAK, T. TÜMER SİVRİ, N. PERVAN AKMAN, A. BERKOL, and Y. EKİCİ, ‘Disease prognosis using machine learning algorithms based on new clinical dataset’, *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, vol. 65, no. 1, pp. 52–68, Jun. 2023, doi: 10.33769/aupse.1215962.

[16] A. Mujumdar and V. Vaidehi, ‘Diabetes Prediction using Machine Learning Algorithms’, in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.

[17] S. Dhabarde, R. Mahajan, S. Mishra, S. Chaudhari, S. Manelu, and N. S. Shelke, ‘DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS’, [Online]. Available: www.irjmets.com

[18] I. D. Mienye, Y. Sun, and Z. Wang, ‘An improved ensemble learning approach for the prediction of heart disease risk’, *Inform Med Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100402.

[19] RVS Technical Campus and Institute of Electrical and Electronics Engineers, ‘Proceedings of the 5th International Conference on Inventive Computation Technologies (ICICT 2020) : 26-28, February 2020’.

[20] Z. Yu, K. Wang, Z. Wan, S. Xie, and Z. Lv, ‘Popular deep learning algorithms for disease prediction: a review’, *Cluster Comput*, vol. 26, no. 2, pp. 1231–1251, Apr. 2023, doi: 10.1007/s10586-022-03707-y.