EasyChair Preprint

№ 3580

Medical Data Analysis Using Machine Learning with KNN

Sabyasachi Mohanty, Astha Mishra and Ankur Saxena

June 9, 2020

MEDICAL DATA ANALYSIS USING MACHINE LEARNING WITH KNN

Sabyasachi Mohanty
Amity University, Uttar Pradesh, India
sabya.mohanty12@gmail.com

Astha Mishra
Amity University, Uttar Pradesh, India
mastha01@gmail.com

Ankur Saxena*
Amity University, Uttar Pradesh, India
asaxena1@amity.edu

Abstract. Machine Learning has been used to develop diagnostic tools in the field of medicine for decades. Huge progress has been made in this area, however, a lot more work has yet to be done in order to make it more pertinent for real-time application in our day-to-day life. As a part of Data Mining, ML learns from previously fed data to classify and cluster relevant information. Hence, the main problems arise due to variations in the big data in the individuals and huge amounts of unorganised datasets. We have used ML to figure out various patterns in our dataset and to calculate the accuracy of this data, with hope that this serves as a stepping stone towards developing tools that can help in medical diagnosis/treatment in future. Creating an efficient diagnostic tool will help improve healthcare to a great extent. We have used a mixed dataset where an individual with any severe illness in early stages or individuals who are further along, are both present. We use libraries like seaborn to construct a detailed map of the data. The fundamental factors considered in this dataset are age, gender, region of stay and Blood groups. The main goal is to compare different data to each other and locate patterns within.

Keywords:*medical diagnosis, seaborn, matplotlib, data mining, KNN*

1 Introduction

Machine learning has helped make huge strides in the fields of science and technology, including medical data processing and significant impact on life science medical research. Few highlights include, the recent advances that have been made in the development of machine learning pipelines for statistical bioinformatics and their deployment in clinical diagnosis, prognosis and drug development [1]. Machine learning algorithms can also be trained to screen complications on medical imaging data [2].

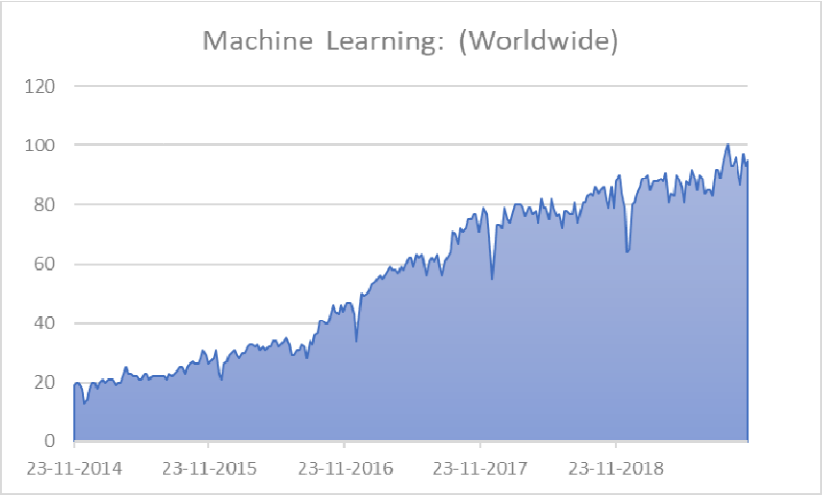

Fig.1. changes in people's interest in Machine Learning over a period 5 years

We obtained this data using google trends which reflect upon the interest people have shown in the field of machine learning since 2014. It is based on the web searches made over this period which a good source to reflect upon the popularity of any kind of entity in this digital age[3]. From 2014 to 2019 there has been consistent rise by huge proportions that shows how vast applications of ML are being realised and discovered by more and more people[4].

Machine learning has gradually spread across several areas within the medical industry with the complete potential to revolutionise the whole industry[5]. Until a few years ago, the medicine solely was dependent on heuristic approaches, the knowledge is gathered through experiences and self-learning, crucial in healthcare environment [6]. The increasing amount of data or the big data is the node for the application of machine learning[7]. ML is a platform that can skim information from numerous sources into an integrated system that can help in decision making processes even for professionals[8].

1.1 Artificial Intelligence

The focus of artificial intelligence has been hugely drawn towards the improvisation of healthcare since the 1960s. In addition to building databases which store medical data such as the patient data, research libraries, administrative and financial systems, the research focus for Artificial Intelligence is innovating techniques for better medical diagnosis[9]. For example, PubMed is a service of the US national library of medicine that includes over 16 million citations from journals for biomedical articles dated back to the 1950s.

1.2 Medical diagnosis

It analyses the structured data sets such as images, genetic and the EP data. In the clinical applications the ML procedures attempt to patient's traits, or infer the probability of the disease results. In the process of molecular drug discovery and manufacturing of drugs, machine learning can be used for precision medicine, next generation sequencing, nano-medicine etc[10]. For better treatments, we are aiming towards the development of improvised algorithms, for example, using the existing treatment methods, say, cancer precision treatment, with the machine learning technologies[11]. Machine learning models have been trained to screen patients. Screening models or the algorithms have already being started for identifying tumours, diabetes, heart diseases, skin cancer etc. The algorithms and ML models should be of high precision and high sensitivity for the best evaluation and diagnosis of the diseases or ailments[12].

Machine Learning tools can be put to various kinds of uses[13]. The following figure 2.0 shows a heat map, that has been used to analyse the Air Quality Index (AQI) of the entire city of Delhi over a month. This data analysis has been performed by a renowned media company news channel, India Today by their Data Intelligence Unit on pollution statistics provided by CPCB[14]. CPCB is Central Pollution Control Board, statutory organisation under the Ministry of Environment, Forest and Climate Change. Therefore, this is publicly available data, which could be easily ignored if not for the processing that India Today did on that impact-less statistical data[15].

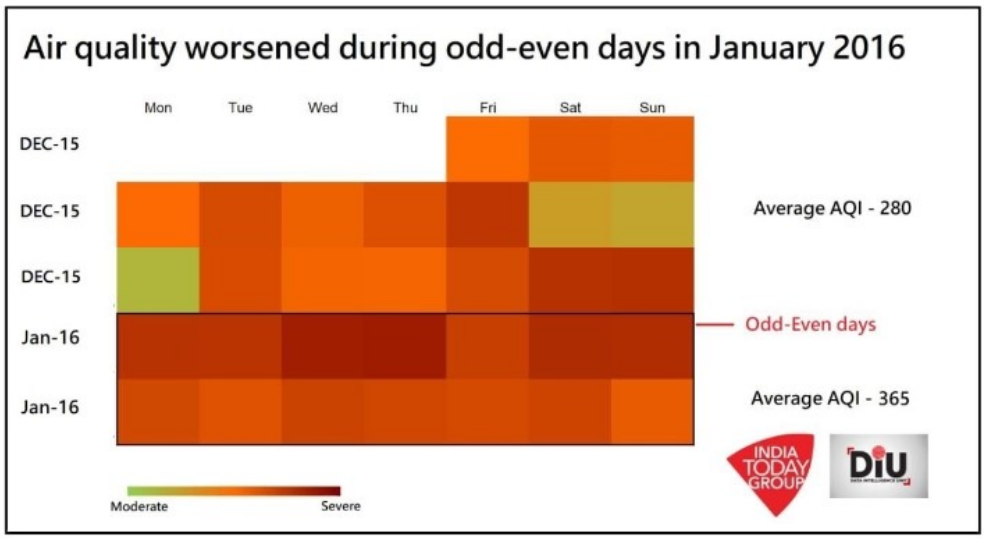

Fig. 2.0 Air Quality Index heat graph for January of 2016 (statistics during odd-even scheme implementation)

One glance at the heat map gives enough information regarding the state of air quality in the city[16]. The dark shades in the odd-even weeks show that air was

at it worst during this period, with average AQI of 365. We are able to analysis the impact of a very popular government scheme without having to read and compare hundreds of numeric values of index[17].

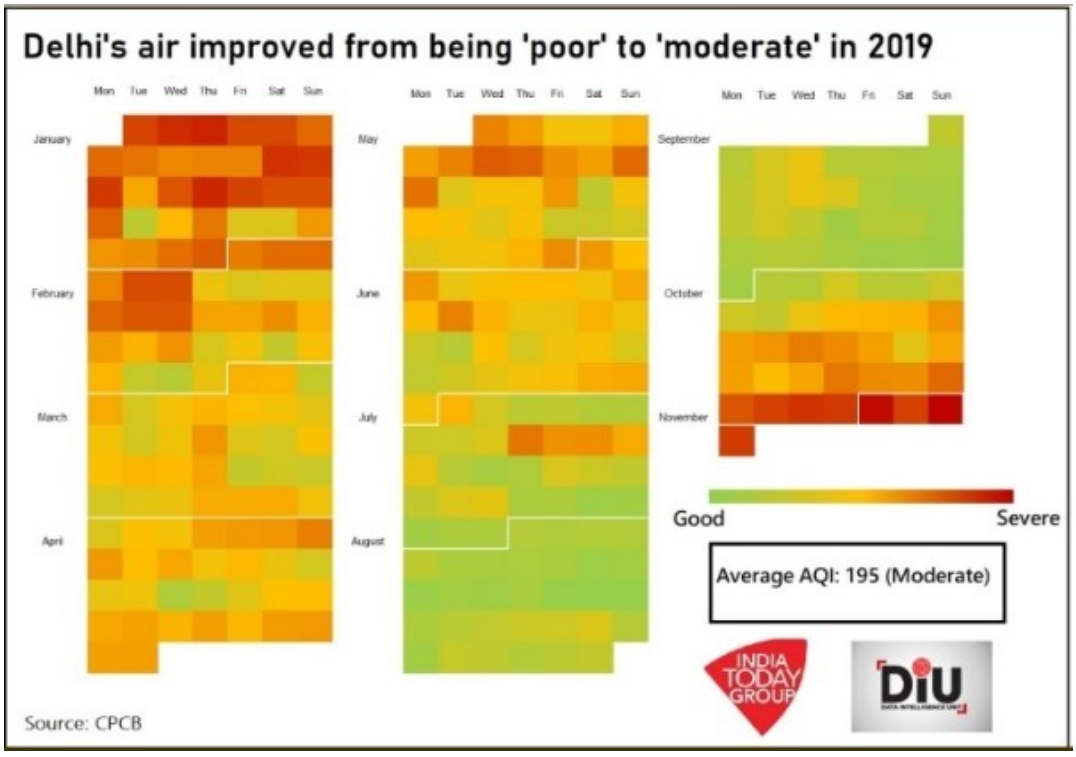

Fig. 3.0AQI heat graph calendar for year of 2019

Use of ML tools in this example has been to analyse large scale Medically and Environmentally relevant data for an area of 1,484 km² with a population of 1.9 crores. Its utility is indeed limitless. The figure 3.0 is a step ahead, it's a pollution Calendar for the year of 2019[18][19].

2 Methodology

2.1 About the Dataset

We have collected this dataset through the means of a google form that we circulated among our college mates and friends. That is why, maximum dataset has the medical information of individuals from the age groups 16 to 30.

Upon receiving complete responses, further processing of dataset involved calculation of BMI from the height and weight data of the individuals, changing certain column entries like Medical History, Symptom Diagnosis, etc. to Boolean format, along with grouping age of individuals into age groups of two years coupled in one group. The dataset processing was instrumental to correct, unambiguous presentation and seamless execution of ML tools on the data.

2.2 Environment Setup

Anaconda was installed to get the work started , as it makes the process of installing libraries seamless, which is used with Python version 3.7 .

We used Jupyter notebook as our IDE because it's one of the gold standard ide for machine learning as it is user friendly and has simple interface. It was most appropriate for our work as it displays the graphs and the data clearly .

2.3 Starting

We used the most popular machine learning libraries of python like Sklearn in our work. The Data was used in CSV (comma separated values) format.

Before starting with the analysis we need to import the libraries and its dependencies. The libraries imported had all things for data analysis machine learning and data visualisation.Pandas, Numpy, matplotlib,seaborn are few major libraries.

The dataset looks like:

```
print (df.head())
```

	AGE	REGION	GENDER	BMI	BLOOD GROUP	medical history	SYMPTOMS	MEDICATIOS	DIAGNOSIS
0	18-19	Faridabad	Male	25-29.9	AB+	0	0	0	0
1	20-21	Noida	Female	18.5-24.9	A+	0	0	0	0
2	18-19	Delhi	Female	<18.5	O+	0	0	0	0
3	18-19	Noida	Male	25-29.9	O+	0	0	0	0
4	20-21	Delhi	Female	18.5-24.9	B+	0	0	0	0

Fig. 4. .head() function of pandas shows the first 5 rows of the dataset

The complete process can be summarised as:

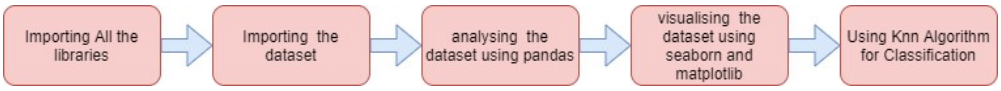

Fig. 5. The workflow

Upon installation of jupyter notebook, an integrated development environment, on our desktop, we used the preinstalled libraries on the software for further editing, like, numpy for mathematical operations, seaborn, sklearn, pandas, matplotlib. Then used pandas library for importing our dataset onto jupyter. We performed

data visualisation using these library functions to helps us with the data analysis process. Then, we used KNN algorithm to classify the data.

3 Results and Discussion

3.1 Data relation with respect to gender using pairplot function of seaborn library:

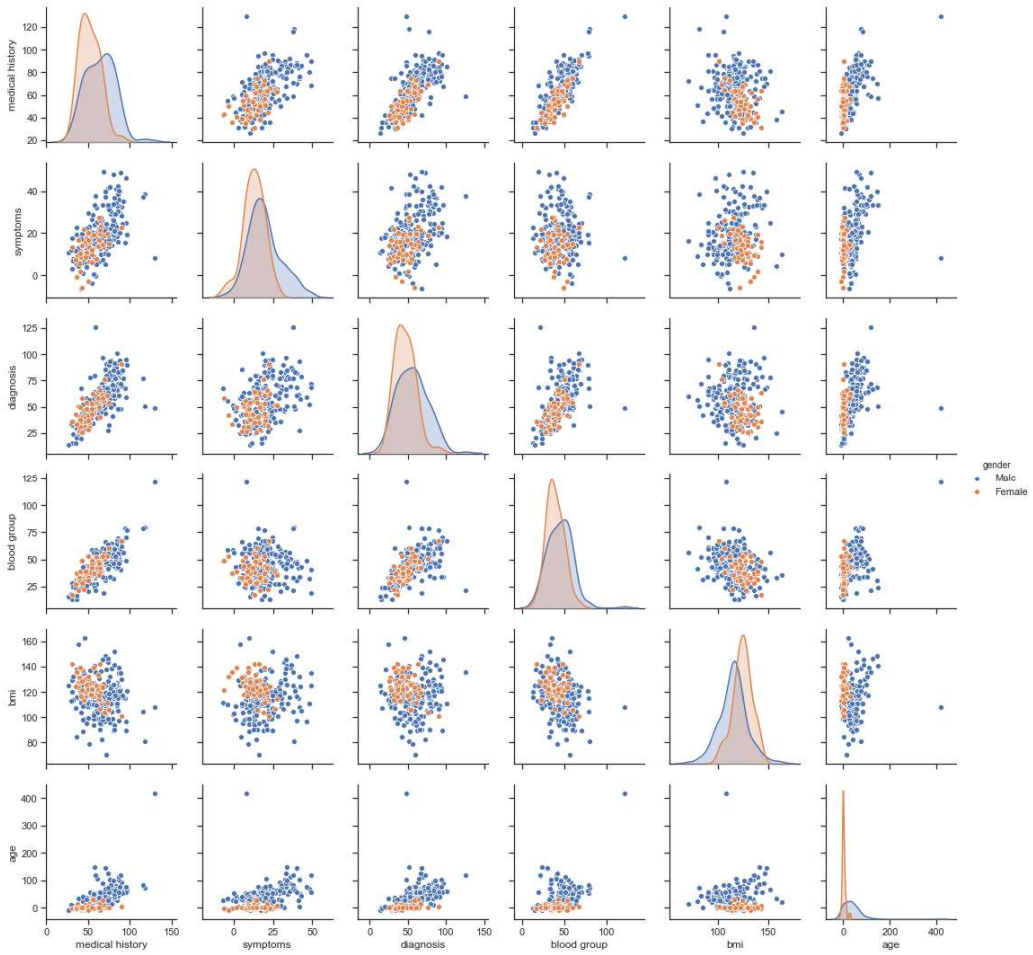

Fig. 6. .pairplot was used to plot the histogram to show relations

The above plot shows the analysis of different parameters of the dataset withrespect to the gender in terms of male and female. The red dots show the female and the blue ones represent the males.We can comprehend various patters in these clusters of points plots against the two axes.

3.2 The few individual parameters of dataset in form of graphs or histograms which is crucial for data analysis :

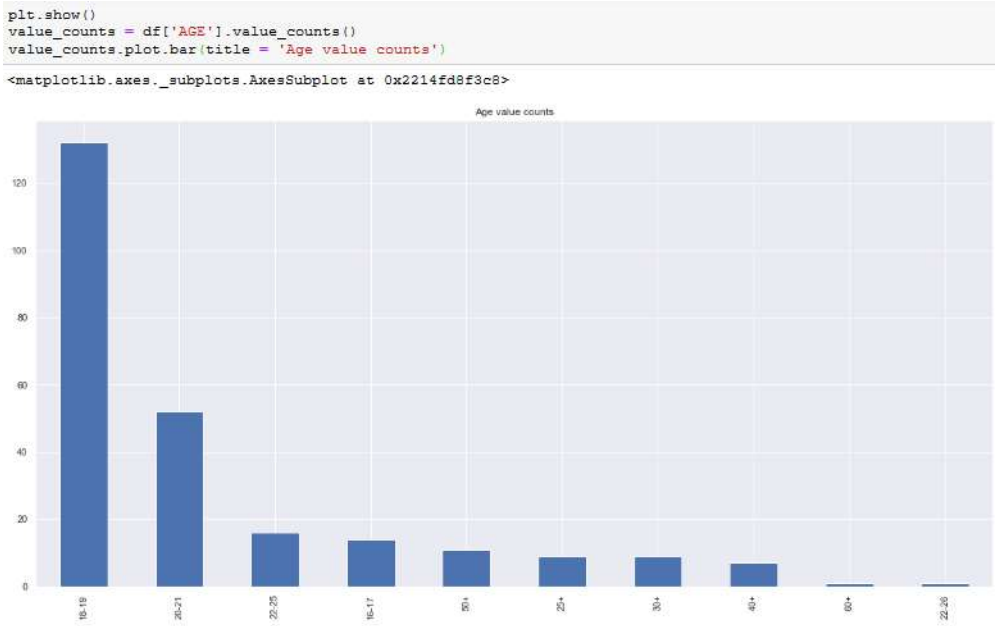

Fig.7. Individual parameter study of age groups in form of histogram

The above histogram gives an overview of the age of the participants. It shows the data has a broad group of participants between the ages of 18 to 21 years as compared to elder groups. This is because since the survey was done a higher educational institute, with majority population of young participants. This helps us prediction the kind of illnesses that could be common in this group among the individuals and what we can expect in general in terms of medical histories.

```
plt.show()
value_counts_1 = df['BLOOD GROUP'].value_counts()
value_counts_1.plot.bar(title = 'Blood group counts')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x221566b7eb8>
```

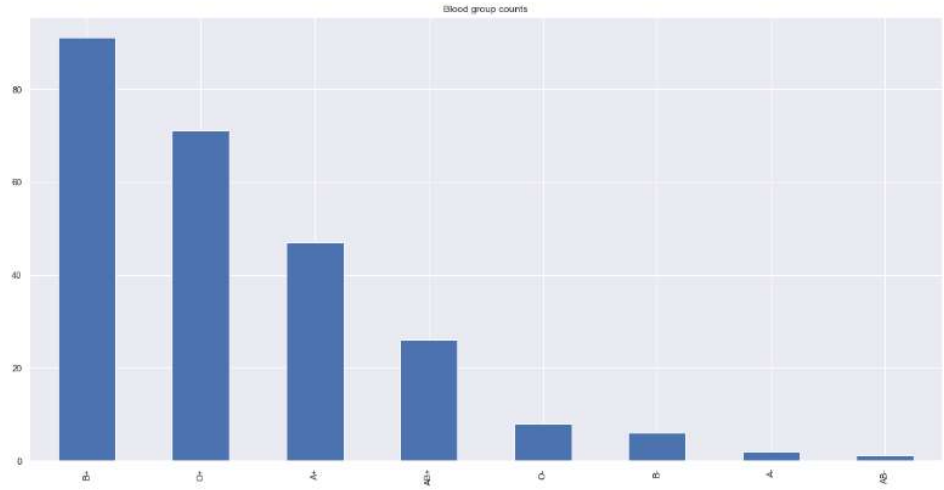

Fig.8. Individual parameter study of blood groups in form of histogram

The above histogram shows that highest number of individuals have B + ve blood group, it is also the most common blood among people of Indian Subcontinent. This definitely speaks well in the accuracy of thisdataset.

```
plt.show()
value_counts_3 = df['GENDER'].value_counts()
value_counts_3.plot.bar(title = 'GENDER')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x22164fe7b70>
```

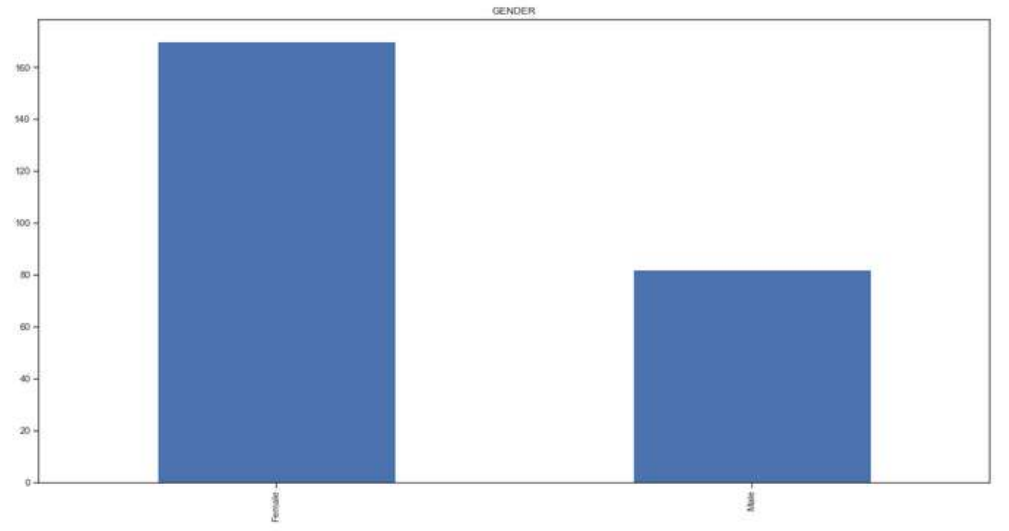

Fig.9. Individual parameter study of blood groups in form of histogram

The survey form circulated through electronic medium, on messaging apps, etc. received significant responses from the female individuals. This can also indicate greater medical and physiological awareness among female participants.

3.3 Comparison of two parameters together:

```
j_df = pd.DataFrame()
j_df['Male'] = df[df['GENDER'] == 'Male']['MEDICATIOS'].value_counts()
j_df['Female'] = df[df['GENDER'] == 'Female']['MEDICATIOS'].value_counts()
j_df.plot.bar(title = 'Gender and medications')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x221567b7940>
```

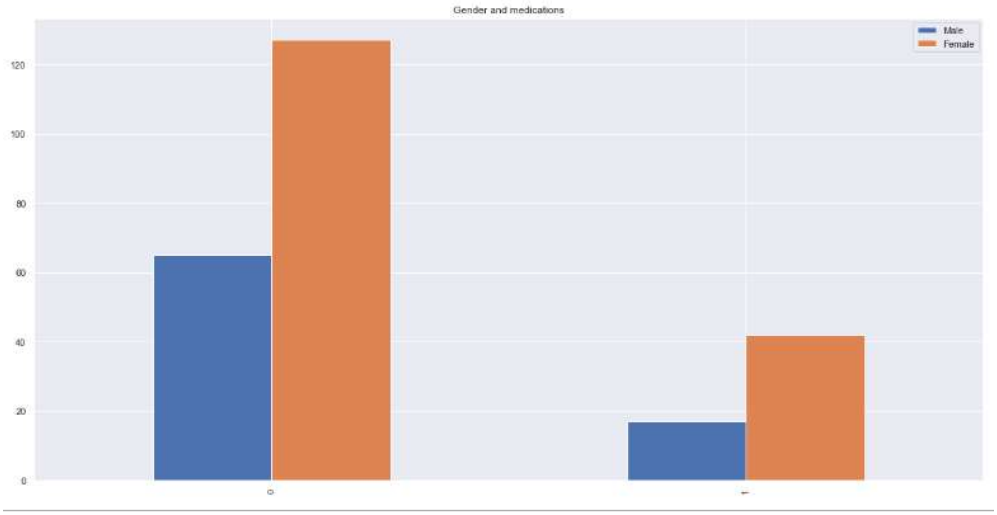

Fig.10. Parameter study of Gender and Medications in form of histogram

The above histogram shows that females took more medications as compared to the male individuals. The above graph proves the fact that females usually consume more medicines and get ill more frequently as compared to men.

```
j_df = pd.DataFrame()
j_df['Male'] = df[df['GENDER'] == 'Male']['BLOOD GROUP'].value_counts()
j_df['Female'] = df[df['GENDER'] == 'Female']['BLOOD GROUP'].value_counts()
j_df.plot.bar(title = 'Gender and Blood Group')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x221567ef2e8>
```

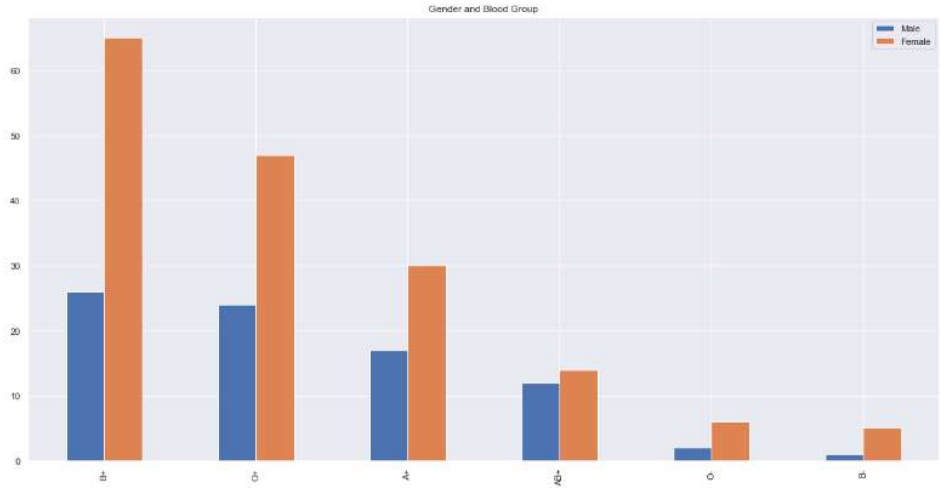

Fig.11. Parameter study of Gender and Blood Group in form of histogram

The above histogram shows that females took more medications as compared to the male individuals. The above graph proves the fact that females usually consume more medicines and get ill more frequently as compared to men.

3.4 For better analysis we did a 1 to 1 comparison of data:

Fig.12. The graph shows the inter relationship of one parameter with each other by a float value.

The heat map gives information regarding the type of data collected by the survey. The dark shade of green shows the right number which had the standard type of data obtained during the data collection. And , thus, lighter shades indicate non-standard data that had to further processed before visualisation and analysis. We did this analysis of the impact factor without having to read and compare hundreds of numeric values of the dataset.

3.5 Finding the K nearest neighbour(KNN) algorithm:

```
plt.scatter(A.medical_history, A.bmi, color = "purple", label = "Female",alpha = 1.0)
plt.scatter(N.medical_history, N.bmi, color = "orange", label = "Male", alpha = 1.0)
plt.xlabel("Medical History")
plt.ylabel("BMI")
plt.legend()
plt.show()
```

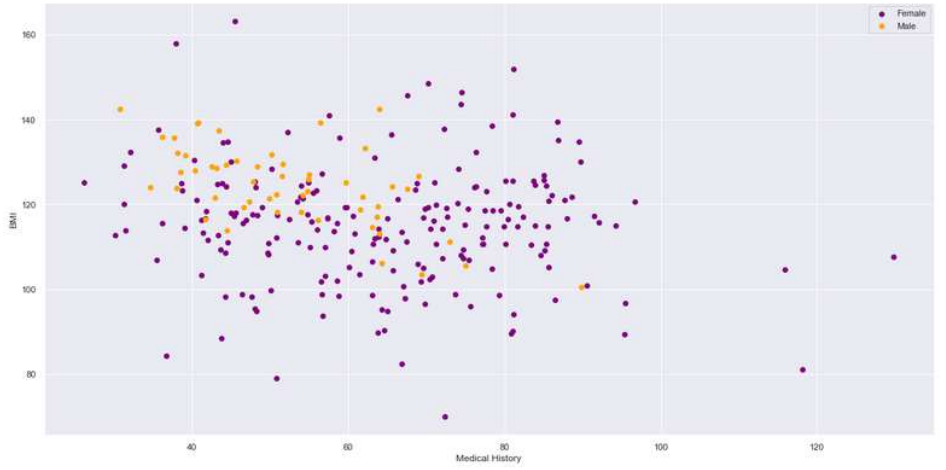

Fig.13. The above histogram plots the number of Males and Females with respect to BMI and Medical History.

This supervised ML algorithm was used to solve classification problems in terms of male and females between the fields BMI and Medical History of the people. Here, similar groups are closer together and the dissimilar ones are relatively farther apart.

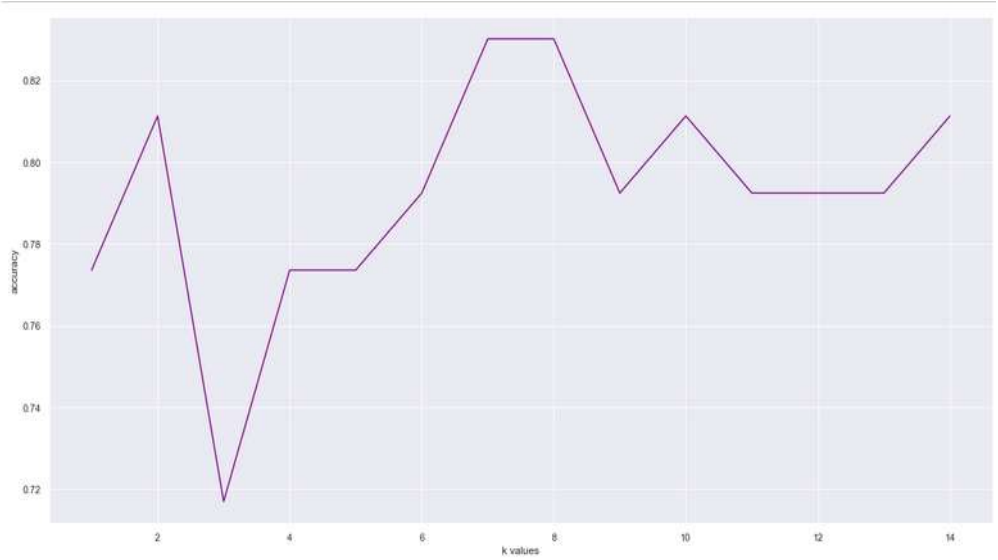

Fig.14. The above graph shows the relation between the accuracy of data with K nearest values**.**

To calculate and present the distance between two corresponding points, we have plotted the above graph for further data analysis in terms of its accuracy (Euclidean Distance or the straight-line distance was used).

4 Conclusion and Future Scope

This model requires better, more robust entries that are accurate and curated. Since, the diagnosis is not specific so it cannot be analysed with just the few parameters as more information is needed to be analysed due to difference in multiple ailments. The data should be 98 % accurate for it to be acceptable in real-time diagnostic tool development. The dataset is required to be trained rigorously to make the analysis more efficient. Also, the future work may involve deep learning and neural network like BERT and other better algorithms after an improvised dataset is formed.

Acknowledgments

We would like to express our deep sense of gratitude towards Amity Institute of Biotechnology and our family, without their support throughout the process this paper would have not been accomplished. We would like to thank Amity Institute of Biotechnology, Amity University for giving us this great opportunity.

References

1. I Sharma , A Agarwal , A Saxena, S Chandra"Development of a better Study Resource for Genetic Disorders through Online Platform"International Journal of Information Systems & Management Science, Vol. 1, No. 2, 2018pp:252-258.
2. S Mohagaonkara, A Rawlani, P Srivastavac, A Saxena"HerbNet: Intelligent Knowledge Discovery in MySQL Database for Acute Ailments"4th International Conference on Computers and Management (ICCM) 2018ELSEVIER-SSRN (ISSN: 1556-5068)pp:161-165.
3. S Shuklaa, A Saxena"Python Based Drug Designing for Alzheimer's Disease" 4th International Conference on Computers and Management (ICCM) 2018ELSEVIER-SSRN (ISSN: 1556-5068),pp:20-24.
4. A Agarwal and A Saxena"Comparing Machine Learning Algorithms to Predict Diabetes inWomen and Visualize Factors Affecting It the Most—A Step Toward Better HealthCare forWomen"International Conference on Innovative Computing and Communications, https://doi.org/10.1007/978-981-15-1286-5_29,2019
5. A Saxena, N Kaushik, A Chaurasia and N Kaushik"Predicting the Outcome of an Election Results Using Sentiment Analysis of Machine Learning"International Conference on Innovative Computing and Communications,https://doi.org/10.1007/978-981-15-1286-5_43,2019
6. A Saxena, S Chandra, A Grover, L Anand and S Jauhari "Genetic Variance

Study in Human on the Basis of Skin/Eye/Hair Pigmentation Using Apache Spark" International Conference on Innovative Computing and Communications, https://doi.org/10.1007/978-981-15-1286-5_31,2019

7. V Vijayan V.,C.Anjali," Prediction and Diagnosis of Diabetes Mellitus - AMachine Learning Approach", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 |Trivandrum.
8. B sarvwar, V Sharma, "Intelligent Naive Bayes Approach to Diagnose Diabetes Type- 2", Special Issue of International Journal of Computer Applications on Issues and Challenges in Networking Intelligence and Computing Technologies, November2012.
9. R Motka, V Parmar, "Diabetes Mellitus Forecast Using Different Data mining Techniques", *IEEE International Conference on Computer and Communication Technology (ICCCT)*, 2013.
10. S. Sapna, A. Tamilarasi, M. Pravin, "Implementation of Genetic Algorithm in Predicting Diabetes", *International journal of computer science issues*, vol. 9, pp.234-240.
11. K Savvas, N. Schizas Christos, "Region based Support Vector Machine Algorithm for Medical Diagnosis on Pima Indian Diabetes Dataset", *IEEE conference on Bioinformatics and Bioengineering*, pp. 139-144,2012.
12. AlJarullah Asma, "Decision discovery for the diagnosis of Type II Diabetes", *IEEE conference on innovations in information technology*, pp. 303-307,2011.
13. D M. Nirmala, Balamurugan S. Appavu alias, U.V. Swathi, "An amalgam KNN to predict Diabetes Mellitus", *IEEE International Conference on Emerging Trends in Computing Communication and Nanotechnology(ICECCN)*, pp. 691-695,2013.
14. U Poonam, H Kaur, P Patil, "Improvement in Prediction Rate and Accuracy of Diabetic Diagnosis System Using Fuzzy Logic Hybrid Combination", *International Conference on Pervasive Computing (ICPC)*, pp. 1-4,2015.
15. S.S Vinod Chandra, S Anand Hareendran, "Artificial intelligence and machine learning" in PHI learning Private Limited, Delhi 110092,2014.
16. R. Bellazzi, B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines", *International Journal of Medical Informatics*, vol. 77, pp. 81-97,2008.
17. A Agarwal, A Saxena, "Malignant Tumor Detection Using Machine Learning through Scikit-learn "International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018,2863-2874

18. Saria S , Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci Transl Med* 2010;2:48ra65.

19. Kale DC , Gong D, Che Zet al. An examination of multivariate time series hashing with applications to health car. In IEEE International Conference on Data Mining (ICDM). 2014. p.260–69.