International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 340-347, 2019, ISSN:-2319–8656

Design of Microstrip Reflective Array Antenna

Meng Li
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Abstract: With the development of communication technology, people have higher and higher requirements for communication quality. Traditional array antennas require complex feed networks and the necessary phase shifters. Conventional reflector antennas are bulky and difficult to manufacture. It is therefore necessary to analyze the traditional characteristics of reflective array antennas to accommodate future adaptability. The microstrip antenna has the advantages of small size, simple structure and small outer shape. Therefore, the working principle and design process of the reflective microstrip array antenna are introduced in detail. A dual-loop antenna operating at $f = 4.5GHz$ was designed, which simplifies the shape of the antenna and achieves a beam pointing of 30°. Compared with similar literature, the new unit antenna has a simple structure, can realize beam orientation without a phase shifter, can work in the low frequency range of 5G, and has high engineering value.

Keywords: Beam orientation; Simple structure; Phase shifter; Microstrip antenna; 5G communication

1. INTRODUCTION

With the rapid development of modern microwave communication, satellite communication and 5G technology, parabolic antennas play an increasingly important role[1-3]. However, due to the modern society's requirements for the flexible operation of communication systems, the disadvantages of the traditional parabolic antennas are cumbersome and bulky[4]. The planar array antenna requires a complicated power distribution feed network, a phase shifter, etc., which can easily increase the transmission loss of the antenna and reduce its transmission efficiency[5]. The microstrip reflective array antenna is operated. The microstrip reflective array combines the advantages of a reflective antenna and a microstrip antenna, and is highly valued for its light weight, small size, low price, and ease of manufacture. Different shapes of reflective elements are discussed in the literature of Dahri, M. Hashim. It can be seen from the discussion of various design and architectural features that the reflective arrays of different structures have different phase shifting ranges[6]. However, it can be seen from the text that

the shape of the reflec Therefore, this paper designs a micro-band unit with a simple double-ring

structure.tion array unit is complicated, which increases the difficulty of production. The structure is simple, can compensate the phase delay of 0-360°, and can realize the beam orientation function of 30° at $f = 4.5GHz$. It also can be used in 5G mobile communication systems and other wireless communication systems in the frequency band, and has high engineering practical value.

2. WORKING PRINCPLE

The microstrip reflection array is mainly composed of two parts, one part is the feed power source and the other part is a reflective medium plate printed with a large number of microstrip arrays.

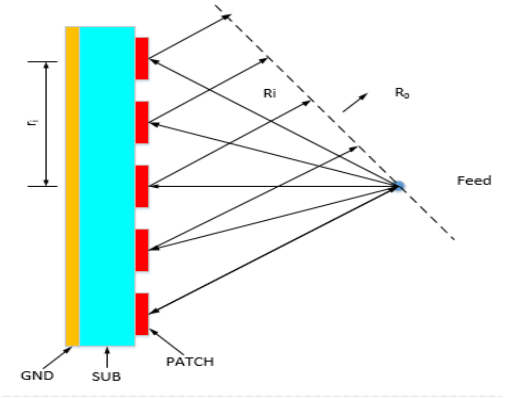


Figure.1 Reflective array antenna working diagram

Working mechanism of microstrip planar reflection antenna: assuming that the planar reflection arrays are all in the far field region of the feed, it can be considered that the electromagnetic waves irradiated to each of the reflection units are plane waves[7]. For a spherical wave, the phase is proportional to the distance between the center of the feed phase and each reflective unit. When the electromagnetic wave radiated by the feed is irradiated from the center of the feed phase to each radiating element on the reflective array, since the transmission distance from the feed to each unit is different, there must be a wave path difference between the respective units, so that each unit is The received incident field has different spatial phase delays, and the size parameters of each unit are reasonably designed according to the phase center of the feed horn and the specified beam pointing, so that it can properly compensate the incident field[8]. From the ray theory, the phase of the ith unit in Figure 1 that needs to be compensated is:

$$\phi_i = 2\pi N + k_0(R_i - r_i.r_0) \quad (N=0,1,2...) \qquad (1)$$

k_0 is the free space wave number、R_i is the phase center of the feed to the position of the i th patch、r_i represents the position vector from the center of the array to the i th patch、r_0 represents the unit vector along the outgoing main beam、$2\pi N$ indicates that the phase compensation period is 2π [8-10].

2.1 Spatial phase delay unit between array elements

The spatial phase delay in a planar array reflective antenna is primarily due to the fact that the array peripheral elements are unequal from the center and thus due to differences in the electric field transmission paths.

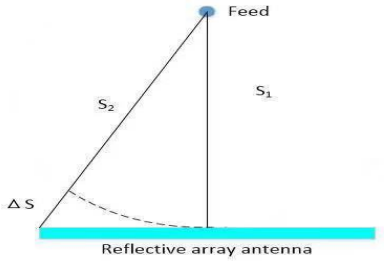


Figure.2 Spatial wave path difference between the units of the reflective array antenna

Therefore, the spatial phase difference that needs to be compensated is:

$$\Delta\varphi = \Delta S \times 2\pi/\lambda \qquad (2)$$

Therefore, in the microstrip array design, $\Delta\varphi$ must be compensated first to achieve the same incident field phase of each unit of the array.

2.2 Analytical method for phase shift characteristics of microstrip reflective array elements

2.2.1 Unit antana model

An isolated unit model, which does not consider the mutual coupling effects of surrounding elements, directly uses plane waves to excite individual isolated elements, and obtains the phase delay generated by electromagnetic waves in the unit according to the phase contrast between the reflected waves and the incident waves. F. Venneri et al. extracted a square

variable size unit antenna using an isolated unit model. This model is fast calculated by computer, but the drawback is that this analysis method only applies to large cell spacing so that mutual coupling can be ignored.

2.2.2 Infinite period model

In the wireless periodic cell model, based on the Floquet theory to simulate the model of the infinite array environment, the influence of the cell spacing on the unit reflection field can be calculated. Therefore, in the unit design, only a single unit model needs to be calculated to complete the wireless large The calculation of the array.

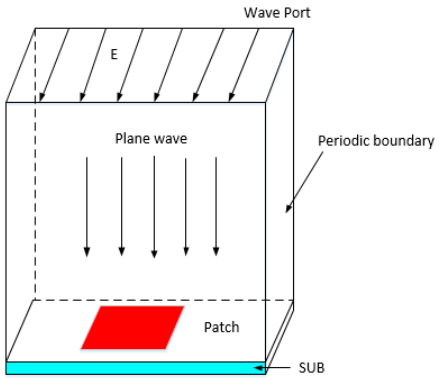


Figure.3 Infinite period model

Similar to the simulation model of a general microstrip antenna, the model mainly consists of three parts, the excitation port, the surrounding periodic boundary and the patch unit. The difference is that the radiation boundary condition in the general model is changed to the periodic boundary condition.

(1) Waveguide Approach-WGA

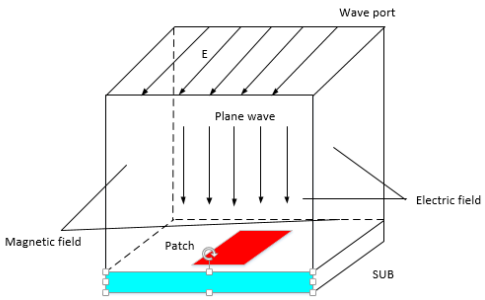


Figure.4 Waveguide simulator model

(2) Master-slave boundary method

The wireless periodic array is simulated by two pairs of master-slave boundaries with a Floquet port. The model excitation port is different from the ordinary wave port by giving two mutually perpendicular electric field excitations to the radiating element on the upper surface of the port, and then porting A layer of PML absorber layer is placed between the master and slave boundaries. This aspect makes up for the simulation flaws that the waveguide simulator can only be used for the normal incidence of plane waves. The master-slave boundary method can illuminate the reflecting unit from any direction, but ensure that the fields on the two master-slave boundaries have the same amplitude and direction.

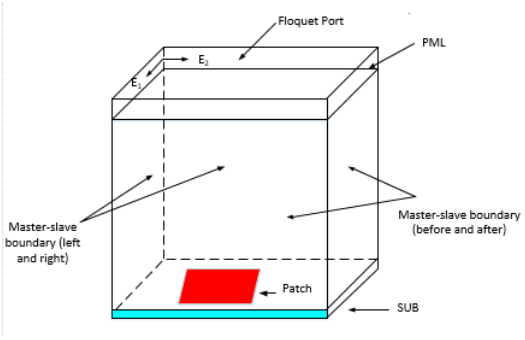


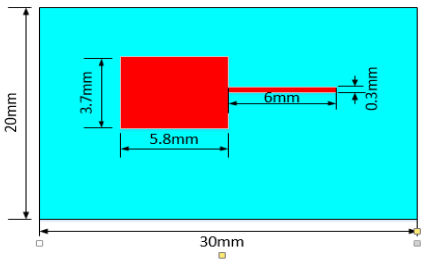
Figure.5 Master-slave boundary method model

2.3 Several typical phase compensation methods

During the analysis, the dielectric material, dielectric thickness, and cell spacing remain unchanged.

2.3.1 Loaded transmission antenna

Each unit of the patch unit in the array has the same shape and size, but the length of the microstrip line connected to it is different, and the required phase shift is adjusted by changing the length of the microstrip line of each patch in the array.



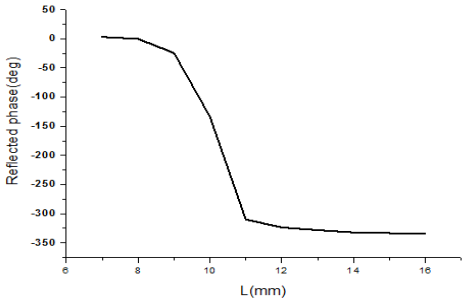
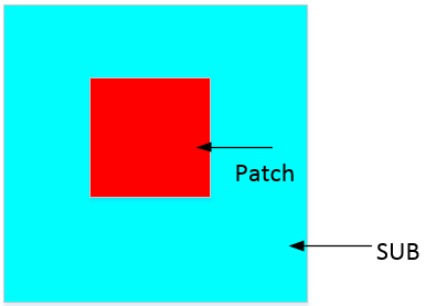


Figure.6 Antenna model and its phase shift range

It is observed from the phase shift curve in the figure that the rectangular patch originating from the load transfer type unit occupies a part of the front surface, so that the phase shift range of the load transfer type unit is very limited (Δ = 332 deg).

2.3.2 Variable size

The shape of the array consists of reflective elements of the same shape but different sizes, and the appropriate amount of phase shift is provided by adjusting the cell patch size.



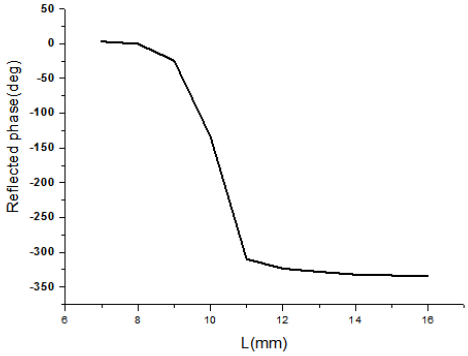
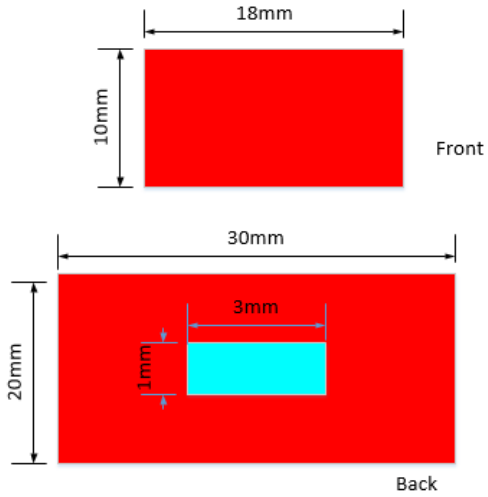


Figure.7 Antenna model and its phase shift range

The phase shifting curve of the square patch unit is close to the "S" shape. It can be seen from the figure that the phase shift range is 327°.

2.3.3 Slotted unit

Each patch unit in the array has the same shape and size, and is slotted on the ground plate on the back side. The size of the slot is determined by the amount of phase shift required.



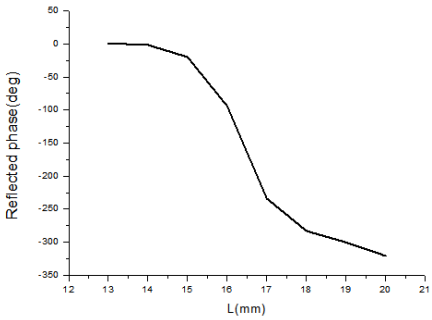


Figure.8 Antenna model and its phase shift range

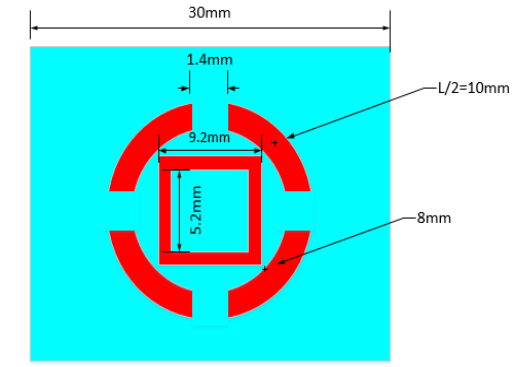
As seen in Figure 8, the range of antenna movement is $\Delta = 363^{\circ}$

3. NEW REFLECTVE ARRAY ANTENNA

3.1 reflection unit structure

A new unit of multi-layer structure is proposed in

the literature [11], and the phase shift range exceeds 450°. The literature [12]. Proposed a windmill type unit with a phase shift range of 700°. However, the two units are complicated in structure and difficult to manufacture. Therefore, this paper designs a new type of unit with double loop structure, the structure is as follows:



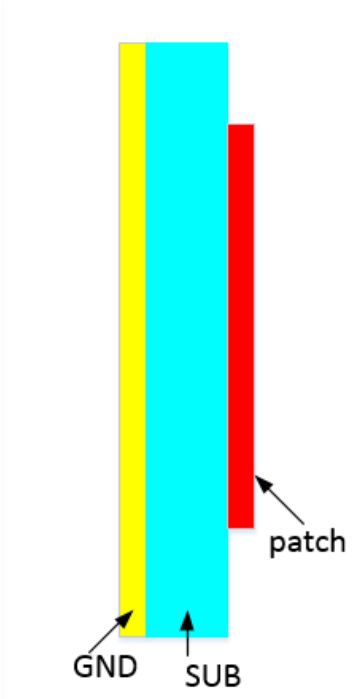


Figure.9 Antenna structure

The unit consists of two annular patches, wherein the inner ring is a square ring, the outer ring is a ring, and there is a gap between the upper and lower sides. Rogers RT/duroid 5880 material with thickness t=2mm, dielectric constant is 2.2. (Related units are marked in the figure).

3.2 Influence of unit structure parameters on phase shift characteristics

（1）Effect of thickness t on phase shifting performance.

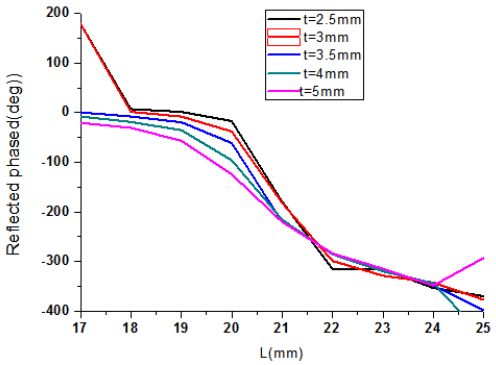


Figure.10 Effect of different parameters d on phase shifting performance

As can be seen from Fig. 10, as the thickness of the substrate increases, the resonance point of the phase shift curve gradually approaches, and the phase shift range is reduced, but still satisfies 360°. Finally take t=4mm.

(2) Effect of parameter d on phase shifting performance

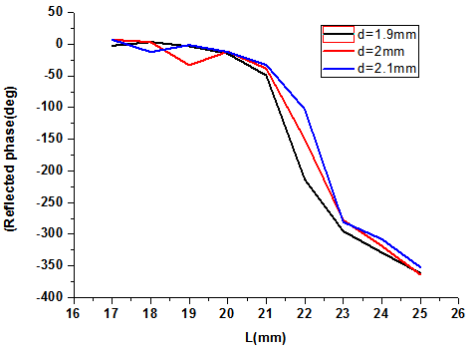


Figure.11 Effect of different parameters d on phase shifting performance

As can be seen from Fig. 11, as the parameter d increases, the outer ring width of the unit antenna increases, the curve becomes steep, and the resonance point distance is zoomed in. Finally, d = 1.9 mm is taken.

(3) Effect of parameter g on phase shift

performance

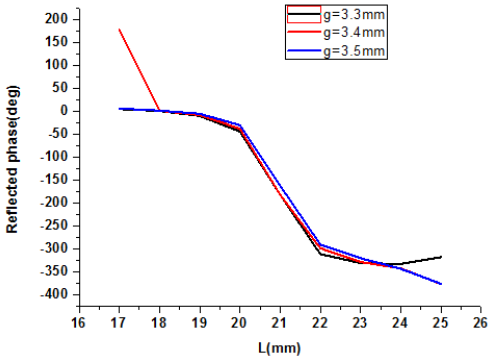


Figure.12 Effect of different parameters g on phase shifting performance

As can be seen from Fig. 12, as the g gradually increases, that is, the interval between the inner and outer rings gradually becomes larger, the phase shift range does not change much. Finally take the height. g=3.3. In summary, the optimized unit parameters are:

Parameters Table 1

Variable	a	t	hs	h	l	d	g
length	30	4	0.035	15	20	1.9	3.3

The optimized curve is shown below:

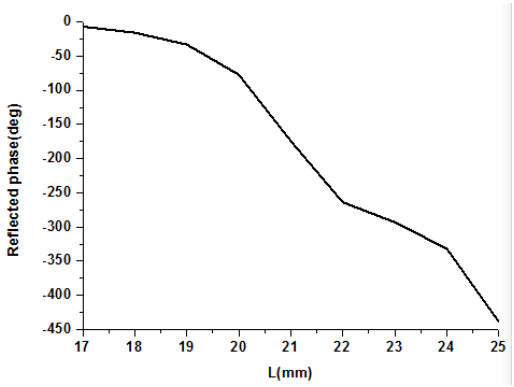


Figure.13 Optimized curve

It can be seen that the phase shift range is $\Delta = 395°$, greater than 360 °, to meet the design requirements.

3.3 Microstrip reflection array design

In this paper, the phase shift performance of the new reflection unit is analyzed and analyzed. The phase shift performance is 0~360°, and the linearity of the phase shift curve is also good. Therefore, the reflecting unit is designed as a reflective array antenna. When the feed is far enough away from the plane array, the feed can be regarded as a plane wave, assuming that the beam points to theta at 30° and the phase delay of the adjacent array unit is:

$$\Delta\phi = -\frac{2\pi f}{c} a\cos\theta \quad (3)$$

f is the working frequency band, c is the speed of light in vacuum, and a is the spacing between adjacent units.

The array adopts 4*4 array mode, the cell spacing is a=12mm, and the dielectric constant is 2.2. The beam is directed to theta=30 degrees, so the phase of the cells in each row and column needs to be compensated:

Parameters Table 2

According to Table 2, the corresponding dimensions of each unit are as follows:

Column / Row Phase	1	2	3	4
1	19	21.75	22.25	23
2	21.75	22.25	23	24.9
3	22.25	23	24.9	21.35
4	23	24.8	21.25	28.1

Make a reflection array based on the dimensional data in Table 3.

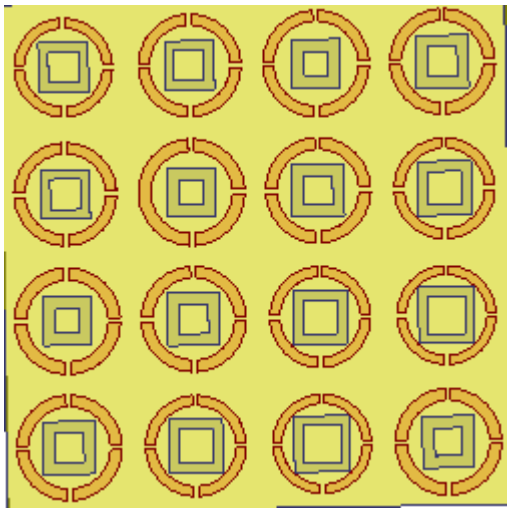


Figure.14 Microstrip reflective array antenna

Simulation and optimization of reflective array antennas using ansoft HFSS software:

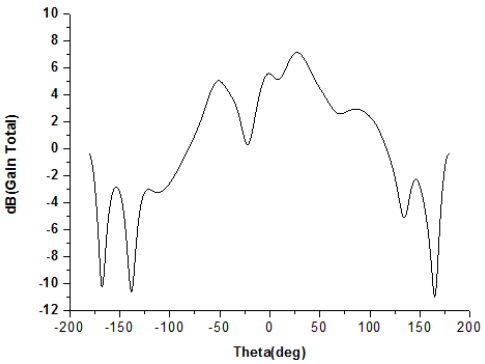


Figure.14 Gain curve of the reflective array

Parameters Table 3

Column / Row Phase	1	2	3	4
1	-33.24	-114.24	-195.24	-276.24
2	-114.24	-195.24	-276.24	-357.24
3	-195.24	-276.24	-357.24	-438.24
4	-276.24	-357.24	-438.24	-519.24

It can be seen from the figure that the pattern of the main beam is in the direction of 30°, which is consistent with the original design, thus verifying that the double-ring unit can achieve beam orientation.

4. CONCLUSION

This paper proposes a microstrip reflective array antenna that can be used in the 5G FR1 band, which compensates for the phase shift of the antenna by changing the structure of the antenna unit. The beam pointing angle is set, and the size of the reflective array unit is calculated. Finally, the main beam of the antenna is accurately oriented to a preset 30° to achieve beam directivity. In the same literature, the unit antenna has a simple structure and is easy to process and design. It can be used in 5G FR1 mobile communication systems and other wireless communication systems, and has high engineering practical value.

5. REFERENCES:

[1] Ta, Son Xuat , H. Choo , and I. Park . "Broadband Printed-Dipole Antenna and Its Arrays for 5G Applications." *IEEE Antennas and Wireless Propagation Letters* PP.99(2017):1-1.

[2] Li, Xichun, et al. "The Future of Mobile Wireless Communication Networks." *International Conference on Communication Software & Networks* 2009.

[3] Sharma, Pankaj . "Evolution of Mobile Wireless Communication Networks-1G to 5G as well as Future Prospective of Next Generation Communication Network." *International Journal of Computer Science & Mobile Computing* 2.8(2013).

[4] Huang, J. "Analysis of a microstrip reflectarray antenna for microspacecraft application." Tda Progress Report 120(1995).

[5] Chang, Zhuang , et al. "A Reconfigurable Graphene Reflectarray for Generation of Vortex THz Waves." IEEE Antennas and Wireless Propagation Letters (2016):1-1.

[6] Dahri, M. Hashim , et al. "A Review of Wideband Reflectarray Antennas for 5G Communication Systems." IEEE AccessPP.99(2017):1-1.

[7] Qin, Pei Yuan , Y. J. Guo , and A. R. Weily . "Broadband Reflectarray Antenna Using Sub-wavelength Elements Based on Double Square Meander-Line Rings." IEEE Transactions on Antennas and Propagation 64.1(2015):1-1.

[8] Chaharmir, M. R. , and J. Shaker . "Design of a broadband, dual-band, large reflectarray using multi open loop elements." Antennas & Propagation Society International SymposiumIEEE, 2010.

[9] Venneri, F. , S. Costanzo , and M. G. Di . "Bandwidth Behavior of Closely Spaced Aperture-Coupled Reflectarrays." International Journal of Antennas and Propagation2012(2012):1-11.

[10] Li, Qin Yi , Y. C. Jiao , and G. Zhao . "A Novel Microstrip Rectangular-Patch/Ring- Combination Reflectarray Element and Its Application." IEEE Antennas & Wireless Propagation Letters 8.4(2009):1119-1122.

[11] José A. Encinar. "Design of two-layer printed reflectarray using patches of variable size." IEEE Transactions on Antennas and Propagation 49.10(2001):1403-1410.

[12] Encinar, J. A. , and J. A. Zornoza . "Broadband design of three-layer printed reflectarrays." IEEE Transactions on Antennas and Propagation 51.7(2003):1662-1664.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 348-352, 2019, ISSN:-2319–8656

The “Promotion” and “Call for Service” Features in the Android-Based Motorcycle Repair Shop Marketplace

Ketut Wahyu Kartika Nugraha
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Made Sukarsa
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Ni Putu Sutramiani
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Abstract: The existence of the motorcycle repair shop business continues to grow, along with the developments of motorcycle riders in Indonesia. However, the majority of riders do not know the existence of the repair shop, especially in the remote location or in the area where they have never visited before. This problem can make that business do not last long. The Motorcycle Repair Shop Information System Application is useful for answering problems related to motorcycle repair shops. "Call for Service" and "Promotion" are two main features of the application which implement E-CRM. The "Call for Service" feature is used to make emergency calls to the nearest repair shop if there is an unexpected situation on the road. The "Promotion" feature is used as a medium to attract as many customers as possible and to increase customer loyalty by providing attractive promotions to the application users. The implementation process uses computers with React Native software, SQLyog, XAMPP, Visual Studio Code and Android smartphones. The Black Box Test in the application reveals that the users can use the “Call for Service” and “Promotion” features from it. The results of data development analysis in the application shows that it only requires a storage space of 73,746 MegaBytes within a year, if there are 25 new data every day.
.

Keywords: E-CRM; mobile application; emergency call; promotion; customer loyalty.

1. INTRODUCTION

The development of motorcycle riders throughout Indonesia has increased by an average of 7.5 million vehicles per year, calculated from 2010 to 2017 [1]. The developing use of motorcycle in Indonesia has opened up opportunities to open repair shop businesses in both small and medium-sized businesses. Motorcycle riders are often faced with difficult situations, for example, sudden flat tire, the engine is not starting, sudden breakdown and so on. The riders are usually not aware of the existence of the nearest small repair shop business. The lack of promotion media for that business also makes it possible that their business will not last long.

The solution created is in the form of an Android-based Motorcycle Repair Shop Information System application aimed at motorcycle riders, and to the owners of the repair shop. The application feature "Call for Service" is intended to overcome the problems experienced by the riders in emergency situations, and the "Promotion" feature will help the repair shop owners to attract as many customers as possible. Both of these features are the implementation of E-CRM in the application in order to maintain good relations between the repair shops and the users of the Android application.

2. LITERATURE REVIEW

A research conducted by Amrapali Dabhade, K.V. Kale and Yogesh Gedam discussed an application that can determine the closest direction to a hospital. The study was used as a reference in the “Call for Service” feature on the Motorcycle Repair Shop application to find the shortest route to a motorcycle rider [2].

A research conducted by Mwangala Mwiya, Jackson Phiri and Gift Lyoko performed a similar study of using GIS (Geographic Information System) technology to report criminal acts to the Zambian police. The research was used in implementing the “Call for Service” feature in the application to provide the location of the user's position [3].

A research conducted by Trinh Le Tan explains the success factor of implementing E-CRM in e-commerce companies. The study was used as a reference for implementing E-CRM on the promotional features contained in the application [4].

3. RESEARCH METHODS

There are four steps in conducting the research. The first one is analyzing the needs from both of the repair shop and the customer. The analysis step is carried out to determine the design of the application, therefore it can answer the needs of both parties. The second step is designing the system workflow. The design of it is done in order to know if the system can perform according to the procedures that have been specified. The third step is to create a system, for both an Android application and a web service which aimed at the admin in managing data. The fourth step is testing the system. The application that have been made will be tested to find out the errors contained in the system, and if there are many errors or malfunctions in the system, a redesign of the workflow will be done to fix the system errors.

3.1 General Overview of the System

The research applications for Android-based Motorcycle Repair Shop Information Systems have a general overview that can be seen in Figure1.

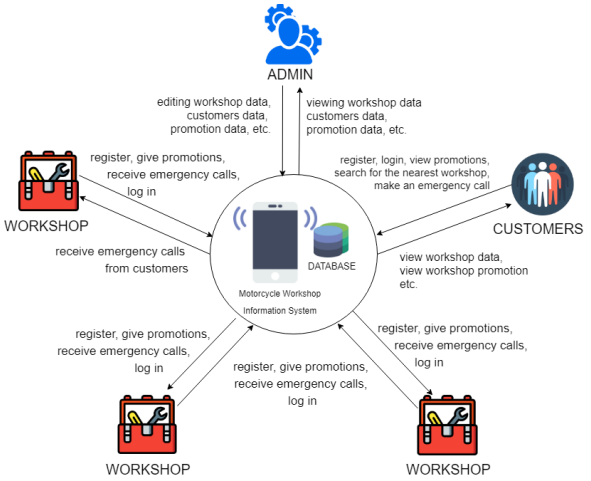


Figure. 1 General Overview of the System

The Motorbike Repair Shop Information System is connected to the database whose data is managed by the admin. These data as if motorcycle repair data, application user data, motorcycle repair shop location data, transaction data and so on. The customer of the application can use it to register as a user, log in to the application, search for the nearest repair shop, view data of all the repair shop, call a repair shop technician using the "Call for Service" feature by using the help of Geographic Information System (GIS), view promotions on the applications and so on. The repair shop can use this application to register their business into the application, login and see the emergency call notifications sent from users, giving promotions and others. The user and the repair shop are connected with the application by using the help of geographic information systems (GIS) mapping.

4. CONCEPTS AND THEORIES

This section contains concepts and theories that support in conducting the research. They are including Android, GIS (Geographic Information System), Google Maps API, Customer Loyalty and E-CRM. It will be discussed as follows.

4.1 Android

Android is a Linux-based operating system used for cellular phones (mobile) such as smartphones and tablet computers (PDAs). It provides an open platform for developers to create their own applications that are used by various mobile devices [5]. Its appearance on March 9th, 2009 introduces an Android version 1.1 and up to the last version 9.0 Pie that has been produced in 2018. Android has been used in everyday life, and moves into all areas of life. It can facilitate transaction activities, for example, in the culinary field, a transaction in a restaurant can now be done only from an Android Smartphone [6]. Game Explore Bali is an application that is engaged in education to educate children about culture in Bali [7].

4.2 GIS (Geographic Information System)

GIS (Geographic Information System) or in Indonesian Language called as *Sistem Informasi Georafis* is an information system that is designed to work by using data that has spatial information (spatial reference). It works by capturing, checking, integrating, manipulating, analyzing, and displaying data that spatially refer to the condition of the earth. The main function of GIS is to conduct spatial data analysis. From the point of view of geographic data processing, GIS is not a new invention. The geographic data processing has been carried out a long time ago by various fields of science, the only difference is that from the use of digital data [8].

4.3 Google Maps API

Google Maps provides an API, it is a provider of digital map services that are popular nowadays. The Google Maps API can be implemented on a web or on an Android / iOS application and provides a map service that can display real images of the earth from satellites, provides a navigation system for travel routes and to find registered places such as business places, recreation areas and so on [9]. The map and navigation system on Google Maps has begun to be developed in the form of augmented reality. The use of augmented reality is intended, therefore the users can improve their driving safety because they can still see the road with a smartphone camera while using maps to navigate routes [10].

4.4 Customer Loyalty

Customer Loyalty or *Loyalitas Pelanggan* is the desire of customers to continue their relationship with a particular company for a long time, it is because the loyal customers are those who buy goods / services of the company from time to time. Loyalty can be interpreted as a customer's desire; a willingness to be a regular customer for a long time; buy and use goods from the selected company and recommending them to friends and colleagues. It is an evidence of the consumers who are always becoming customers, who have the strength and positive attitude towards the company. Each of the customers has a different basis of the loyalty and it depends on their perspective views [11].

4.5 E-CRM

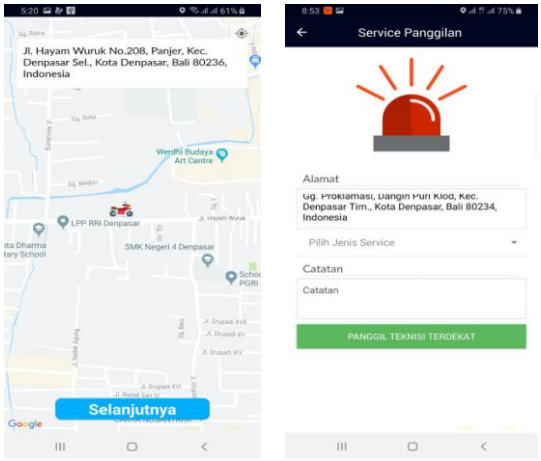
E-CRM is a CRM (Customer Relationship Management) which is implemented electronically by using a web browser, internet, and other electronic media such as e-mail, call centers, and personalization. It is a technique for the companies which is done by online to strengthen the relationship between the company and its customers, where it aims to increase customer satisfaction and gain loyalty from consumers. Also, it has a definition of using digital communication technology to maximize customer sales and encourage the use of online services [12].

5. RESULT AND DISCUSSION

The results and discussion of the Motorcycle Repair Shop Information System application contains the results of testing the system directly, the results of Black Box testing and the results of the analysis of data development. These three results will be discussed as follows.

5.1 System Testing

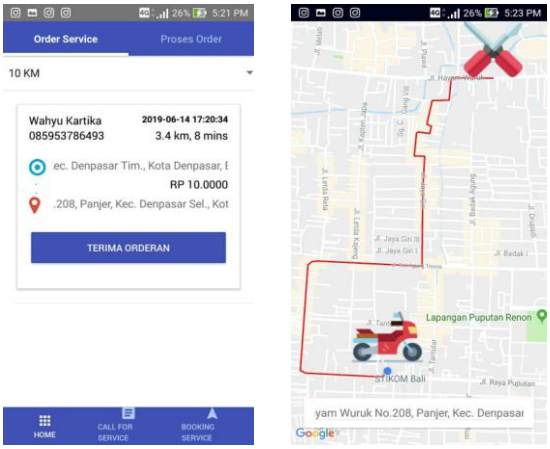
The customer can make emergency calls to nearby repair shops, and the repair shop can also receive emergency calls made by the application user. Testing this system is done directly by using the Motorcycle Repair Shop Information System application. The call from the customer to the nearest repair shop, is displayed in Figure 2.



(a) (b)

Figure. 2 "Call For Service" Feature

Figure 2 shows the step of selecting a customer's location before making an emergency call from the customer's application. This location selection is intended in order to be more accurate towards the customer location. Figure 2(a) is the step of displaying a map in order to select the location of the customers when making emergency calls. In Figure 2(b), the customer is asked to choose one type of the damage and can include notes for the repair shop technician. The display of "Call for Service" from the repair shop application point of view is displayed in Figure 3.



(a) (b)

Figure. 3 "Call for Service" Feature

Figure 3(a) is a display of incoming emergency calls from the repair shop's application. The repair shop can receive the call by pressing the "*TERIMA ORDERAN*" button, or ignore the call if they do not want to receive it. Figure 3(b) is a navigation display of directions to the customer's location who make emergency calls. A research conducted by Yuli Fauziah, Heru Cahya Rustamaji and Rihadina Ramadhan created an application that can predict the arrival of Trans Jogja buses by broadcasting locations to passengers, therefore the estimated arrival time can be predicted [13]. A research conducted by Made Yudha Putra Mahendra, I Nyoman Piarsa and Dwi Putra Githa produced a public complaint application by using the Geographic Information System to record the location of a complaint, and the admin could read all the community complaints and find out the location of it [14]. Both of the studies are used as references to predict the mechanic's arrival time to the customer and to find out the customer's location in the Motorcycle Repair Shop Information System application.

5.2 Black Box Testing Analysis

The black box testing or referred as functional testing, is a testing technique regarding to the function of a system based on a particular test case. The people who perform black box testing do not have direct access to the application source code, but they only focus on the output produced as a response to the input chosen by the examiner and the execution conditions of the system [14]. The table of black box testing can be seen in Table 1.

Table 1. Black Box Testing Analysis

Testing Activity	The Expected Realization	Testing Result	Result
Adding repair shop promotions	The added promotion data successfully appears in the repair shop's application promotion menu	New promotion data has been successfully added and appears in the repair shop's application	[x] Accepted [ ] Rejected
Changing repair shop promotion data	The promotion data successfully changed in the repair shop's application	The repair shop promotion data successfully changed	[x] Accepted [ ] Rejected
Removing repair shop promotion	The promotion data that want to be deleted, successfully deleted in the application promotion menu	The deleted promotion data is disappear from the promotion menu in the repair shop's application	[x] Accepted [ ] Rejected
Looking at a list of repair shops that have promotions from customer's applications	The repair shops that have promotions are marked with a green indicator which means "Promotion"	There is an indicator in green which means "Promotion" at a repair shop that has a promotion	[x] Accepted [ ] Rejected
Spotting promotions from the Promotions menu through the customer's application	An added list of promotion provided by repair shop to their application appears	A list of promotion data provided by the repair shop appears	[x] Accepted [ ] Rejected
Booking a service from certain promotions	The customers successfully book a	The booking service was successfully made, but	[x] Accepted

	service with certain promotions	the promotion calendar system has not functioned properly	[ ] Rejected
Making an emergency call to the nearest technician through the customer's application	The customers can choose their location, fill out a complaint about their vehicles and find the nearest repair shop from their locations.	The customer successfully chooses their location, includes their complaints and makes an emergency call.	[x] Accepted [ ] Rejected
Receiving emergency calls from customers through repair shop's application	The technicians can notice emergency calls from the customers and can receive them.	The technicians successfully see the emergency call from the customers and successfully receive it.	[x] Accepted [ ] Rejected
Reviewing the list of emergency calls from the repair shop's application	The repair shop can see all of the received emergency calls, along with the status of the call.	The repair shop can see all of the received emergency calls, but cannot see the status of that call.	[x] Accepted [ ] Rejected
Navigating the customer's location from the repair shop's application	The technicians can navigate the direction of customer locations through digital maps.	The technicians can navigate the direction of customer's location in a digital map.	[x] Accepted [ ] Rejected
Tracking the location of the technician from the customer's application	The customer can monitor the presence of the technician who receives emergency calls that has been made in real-time.	The customer cannot monitor the presence of the technician.	[x] Accepted [ ] Rejected
Looking at an emergency call transaction history from the customer's application.	The customer can see an emergency call transaction history that has been made, along with the total price charged.	The customer cannot see an emergency call transaction history that has been made.	[x] Accepted [ ] Rejected
Looking at an emergency call transaction history from the repair shop's application.	The repair shop can see an emergency call transaction history that has been received, and show the services that they have performed, along with the total price charged.	The repair shop can see a history list of emergency call transactions that has been received, but cannot see the services that they have performed, along with the total price charged.	[x] Accepted [ ] Rejected

The black box test results in Table 1 indicates that the repair shop can create a new promotion data that will be provided to the customers. They can change the promotion data that already exists in their promotion data. Also, they can delete it in the menu from their application. The customer can see a list of the repair shop that provides promotion from their application, both from the repair shop list menu or when they book a service in the booking menu. They can see various promotion lists that appear in the Promotions menu from their application. In addition, they can directly order services based on the certain promotions on the Promotion menu. The customer can make an emergency call to the nearest repair shop technician and track them. The transaction history of an emergency call also can be seen by the customer. The repair shop can receive an emergency call from the customer and navigate the direction to their location through the digital maps. They also can see the status and the history of the received emergency call.

5.3 Data Growth Analysis

This section will tell an explanation of the estimated system data storage space requirements in the database. That estimations are used to predict the database's ability to store data. The analysis is done by calculating the type of storage space requirements based on the data of each table which is required on the system. The tables in the Motorcycle Repair Shop Information System database are classified into 2 groups, such as the Transaction Table and the Master Table. The analysis of data growth from both of the groups can be seen in Table 2.

Table 2. Data Growth Analysis

		Master Table	Transaction Table
The Number of Table		4	8
1 Row Data Storage Requirement (*Kilo Bytes*)		2	8,849
The Amount of Data Per Day		25	25
The Estimated Storage Space Requirement (*Kilo Bytes*)	1 Day	55	221,225
	30 Days	1.652,25	6.636,75
	365 Days	20.102,39	73.746,133

Table 2 shows an analysis of data growth from the Master Table and Transaction Table groups, with the assumption there are 25 data per day. The results reveal that the Master Table requires a storage space of 55 kilobytes for a day; 1,652.25 kilobytes for 30 days; and 20,102.39 kilobytes for 365 days. In other hand, the Transaction Table requires storage space of 221,225 kilobytes for a day; 6,636.75 kilobytes for 30 days; and 73,746,133 kilobytes for 365 days.

6. CONCLUSION

The Motorcycle Repair Shop Information System Application is an Android-based marketplace application that aims to improve the economic level of repair shop business, and help the riders in everywhere and at any time by implementing E-CRM on the "Call for Service" feature through the application. The Black Box Testing in the application shows that the user can use the "Call For Service" feature, and reveals that the application has successfully implemented E-CRM in that feature in order to make an emergency call. In testing data development analysis, it shows that the Motor Repair Shop Information System application only requires a storage space of 73,746,133 kiloBytes (73,746 MegaBytes) for 365 days, if it is assumed that there are 25 new data per day. In the future, the application can still be developed both in terms of display and new features, such as a "live chat" feature with the mechanics when the customers use the "Call for Service" feature in order to make it easier to communicate with both parties.

REFERENCES

[1] POLRI, "The Number of Motorcycle Developments Based on Its Type, from 1987-2008 (In Indonesian: "Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis tahun 1987-2008)," 2009. [Online]. Available: http://www.bps.go.id/tab_sub/view.php?tabel=1&daftar=1&id_subyek=17&notab=12. [Accessed: 06-May-2019].

[2] A. Dabhade, K. V Kale, and Y. Gedam, "Network Analysis for Finding Shortest Path in Hospital Information System," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 5, no. 7, pp. 618–623, 2015.

[3] M. Mwiya, J. Phiri, and G. Lyoko, "Public Crime Reporting and Monitoring System Model Using GSM and GIS Technologies : A Case of Zambia Police Service," *Int. J. Comput. Sci. Mob. Comput.*, vol. 4, no. 11, pp. 207–226, 2015.

[4] T. Le Tan, "Successful Factors of Implementation Electronic Customer Relationship Management (e-CRM) on E-commerce Company," *Am. J. Softw. Eng. Appl.*, vol. 6, no. 5, p. 121, 2017.

[5] T. Cui, Y. Wu, and Y. Tong, "Exploring ideation and implementation openness in open innovation projects: IT-enabled absorptive capacity perspective," *Inf. Manag.*, vol. 55, no. 5, pp. 576–587, 2018.

[6] I. K. K. Sanjaya, P. W. Buana, and I. M. Sukarsa, "Designing Mobile Transactional Based Restaurant Management," *Int. J. Comput. Eng. Inf. Technol.*, vol. 11, no. 6, pp. 130–136, 2019.

[7] D. P. A. Sanjaya, I. K. A. Purnawan, and N. K. D. Rusjayanthi, "An Introduction to Balinese Cultural Traditions through the Android-Based Game Explore Bali Application (In Indonesian: Pengenalan Tradisi Budaya Bali melalui Aplikasi Game Explore Bali Berbasis Android)," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 7, no. 3, pp. 162–173, 2016.

[8] E. A. Sholarin and J. L. Awange, "Geographical information system (GIS)," in *Environmental Science and Engineering (Subseries: Environmental Science)*, 2015.

[9] A. Rahmi, I. N. Piarsa, and P. W. Buana, "FinDoctor – Interactive Android Clinic Geographical Information System Using Firebase and Google Maps API," *Int. J. New Technol. Res.*, vol. 3, no. 7, pp. 8–12, 2017.

[10] I. N. Piarsa, P. W. Buana, and I. G. A. Mahasadhu, "Android Navigation Application with Location-Based Augmented Reality," *Int. J. Comput. Sci. Issues*, vol. 13, no. 4, 2016.

[11] M. Išoraitė, "Customer Loyalty Theoretical Aspects," *Ecoforum*, vol. 5, no. 2, pp. 292–299, 2016.

[12] A. B. Ramadhan, "The Role Of E-Crm (Electronic Customer Relationship Management) in Improving Service Quality (Study at Harris Hotel & Conventions Malang) (In Indonesian: Peran E-CRM (Electronic Customer Relationship Managemen) dalam Meningkatkan Kualitas Pelayanan (Studi pada Harris Hotel & Conventions Malang))," *J. Adm. Bisnis*, vol. 40, no. 1, pp. 194–198, 2016.

[13] Y. Fauziah, H. C. Rustamaji, and R. P. Ramadhan, "The Implementation of Mobile Crowdsourcing for Estimating Bus Arrival Times Based on Community Information (In Indonesian: Penerapan Mobile Crowdsourching Untuk Estimasi Waktu Kedatangan Bis Berdasarkan Informasi Masyarakat)," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 7, no. 3, p. 150, 2017.

[14] M. Y. P. Mahendra, I. N. Piarsa, and D. Putra Githa, "Geographic Information System of Public Complaint Testing Based On Mobile Web (Public Complaint)," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 9, no. 2, p. 95, 2018.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 353-357, 2019, ISSN:-2319–8656

Optimization of Lift Gas Allocation using Evolutionary Algorithms

Sofía López
University of Barcelona
Spain

Urhan Koç
Istanbul Polytechnic
Turkey

Emma Bakker
Leiden University
Netherlands

Javad Rahmani
Islamic Azad University
Iran

Abstract: In this paper, the particle swarm optimization (PSO) algorithm is proposed to solve the lift gas optimization problem in the crude oil production industry. Two evolutionary algorithms, genetic algorithm (GA) and PSO, are applied to optimize the gas distribution for oil lifting problem for a 6-well and a 56-well site. The performance plots of the gas intakes are estimated through the artificial neural network (ANN) method in MATLAB. Comparing the simulation results using the evolutionary optimization algorithms and the classical methods, proved the better performance and faster convergence of the evolutionary methods over the classical approaches. Moreover, the convergence rate of PSO is 13 times faster than GA's for this problem.

Keywords: particle swarm optimization; crude oil lifting; lift gas allocation; optimization; artificial neural network; genetic algorithm.

1. INTRODUCTION

There exist a wide variety of natural mechanisms to drive crude oil from the underground reservoirs to the surface, including the gas expansion and water pressure mechanisms. When the natural energies to produce crude oil from a well is not sufficient, the artificial lift procedures are used to accomplish the oil production process. In general, the artificial lift processes are divided into two main categories; gas-based lift process and pump-based lift process [1-8]. Gas-based lift technology is known as an efficient and economical procedure in the oil production industry. In a gas-based lift process, the optimal rate for the gas injection is determined such that it can compensate for the hydro-static pressure drop and frictional pressure drop in the well [9]. The optimum injection rate is important, mainly because of the operating constraints related to the available gas intake.

One of the very first studies on gas allocation optimization was conducted by Redden et al. in 1974 [10]. Authors in [10] have optimized the gas distribution among 30 wells in Venezuela. Their approach was based on the good laboratory practices (GLP) diagrams, and the optimization criterion was the higher profit rate. Their proposed strategy did not consider any optimization constraints; i.e., they assumed the unlimited amount of gas is available. A similar study is conducted by Mayhill in 1974 [11]. In 1981, Kanu and his colleagues introduced a parameter called the economic slope, which was a measure of the economic efficiency in a gas-based lift process. In their proposed approach, the optimal gas allocation was analyzed with and without constraints; e.g., with limited and unlimited gas intake [12]. In a further study, Nishikiori et al. developed a strategy based on the economic slope parameters, in which the optimum amount of gas injection was determined through a pseudo-Newtonian method [13]. In [14], authors optimized the controller tuning process using the particle swarm optimization. The objective of the optimization problem in their approach was to maximize the production rate. They also utilized GLP diagrams in their method. In another study, [15] developed a distributed algorithm to optimize the energy allocation in a building environment [15]. In [16], the rate of lift gas injection is determined based on the net present value (NPV). From their study, it is realized that the maximum profit from the production does not necessarily occur at maximum production. Authors also proved that the oil price is an important parameter in the optimization process, and an appropriate optimization scenario should be picked considering the oil price rate. However, the authors did not provide a well-designed model for their strategy. [17] applied the control theory principles to optimize the lift gas distribution; their approach was a cascaded control strategy. [18] developed an algorithm based on ant colony algorithm (known as continuous ant colony optimization, or CACO) to solve the gas allocation problem.

In this paper, the optimum amount of lift gas is distributed over a set of wells based on an evolutionary optimization algorithm. It is the first time that the particle swarm optimization (PSO) algorithm is used for finding the optimal gas injection rate for oil lift process. Worth mentioning that PSO algorithm is known to be more efficient and faster in solving such optimization problems, compared to the similar evolutionary algorithms such as genetic algorithm (GA). Moreover, in this study, the artificial neural network (ANN) method is utilized to estimate the performance plots of the gas-based lift process.

The rest of the paper is organized as follows. The next section explains the two evolutionary algorithms; genetic algorithm (GA) and particle swarm optimization (PSO). Section 3 describes the PSO algorithm challenges. The proposed strategy is shown in section 4. Section 5 includes the simulation results. The work finishes with the conclusions in section 6.

2. EVOLUTIONARY ALGORITHMS

In this section, the genetic optimization algorithm (GA) and the particle swarm optimization (PSO) algorithm are explained in detail.

2.1 Genetic algorithm

Genetic algorithm is one of the most important meta-heuristic algorithms, first introduced by Holland in 1975 [19]. Genetic algorithm is a type of evolutionary algorithm, which is commonly used in artificial intelligence (AI) and computing. The genetic algorithm applies a set of solutions to the optimization problem in each generation. The selection process chooses the individuals with the best fitness; these individuals mutate and reproduce new genes [20-26]. Therefore, the best optimum solutions are attained through mimicking the natural process genes mutation, selection, and reproduction. In the genetic algorithm, the final goal of selections and mutations is to maximize the fitness or minimize the costs of each individual. The genes adapt themselves to the environmental conditions such that they survive or mutate with genes with higher fitness. The crossover operator is used to produce new offsprings from every two parents.

2.2 Particle swarm optimization algorithm

Extensive studies have investigated the social behavior of various types of creatures; such as birds flock, school of whales, fish, sharks, etc. The particle swarm optimization (PSO) algorithm is a meta-heuristic computational method that mimics the social behavior of animal swarms. PSO optimizes problem by improving the candidate solution iteratively. The algorithm was first introduced by Kennedy and Eberhart in 1995 [27]. Swarm intelligence is the collective behavior of self-organized systems. The algorithms in artificial intelligence (AI) follow a hierarchy directly or indirectly. In PSO algorithm, two main parameters are being updated in each iteration; velocity term and position term. The particle's velocity and position are updated through the following equations, respectively.

$$v_i(t+1) = wv_i(t) + c_1 r_1(y_i(t) - x_i(t)) + c_2 r_2(\hat{y}(t) - x_i(t)) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1)$$
$$x_i(t) \in U[x_{min},\ x_{max}] \quad (2)$$

where vi(t) and xi(t) denote the velocity and position of particles at time t. y and parameters represent the personal best solution of the particle and the global best solution, respectively. r1 and r2 are the random vectors with uniform distribution in the [0,1] interval. w, c1, and c2 are the inertia coefficient, personal learning coefficient, and collective learning coefficient, respectively.

Beside the velocity and position updates, the personal best and global best parameters should also be updated in a standard PSO algorithm.

$$y_i(t+1) = y_i(t) \quad for f(x_i(t+1)) \geq f(y_i(t))$$
$$y_i(t+1) = x_i(t+1) \quad for f(x_i(t+1)) \leq f(y_i(t))$$
$$\hat{y}(t) = y_0, y_1, ..., y_z = min f(y_0(t)), f(y_1(t)), ..., f(y_z(t)) \quad (3)$$

The PSO algorithm is as follows.

- For each particle $i \in 1, .., s$, initialize th position x_i and velocity v_i randomly.
- Set $y_i = x_i$.
- For each particle i, evaluate the fitness function $f(x_i)$.
- For each particle i, update y_i and $\hat{y}_i$ from (3).
- For each dimension $j \in 1, ..., N_d$, update the velocity from:

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_{1,j}(y_{i,j}(t) - x_{i,j}(t)) + c_2 r_{2,j}(\hat{y}_j(t) - x_{i,j}(t)) \quad (4)$$

- Apply the position update to each particle.
- Stop the algorithm when the convergence criteria is met, otherwise, go to step 3.

3. PARTICLE SWARM OPTIMIZATION ALGORITHM CHALLENGES

The particle swarm optimization algorithm has several drawbacks and disadvantages. PSO can easily fall into the local optimum points in high-dimensional optimization problems. Although PSO is faster compared to similar evolutionary algorithms, its convergence rate does not enhance with a higher number of iterations. The prominent reason is that in this algorithm, particles converge to the point with the personal best and global best solution. To address this issue, the inertia weight w is used to modify the algorithm [28]. Another main drawback in this algorithm is that the quality of solutions is very much dependent to the weighting coefficients and algorithm parameters [29]. Therefore, we should try to tune the PSO parameters in the best way.

4. PROPOSED STRATEGY

In order to define the optimization problem, we first need to estimate the performance diagrams of the wells with different levels of gas injections. The artificial neural network (ANN) algorithm is utilized in this step to attain the (good laboratory practices) GLP-based performance diagrams. The training model is then used as the fitness function in the optimization process. Once the convergence criteria are met, the algorithm stops. The PSO algorithm is simulated in MATLAB environment. The advantages of coding in MATLAB include:

- The programmer can do the code testing, implementing, visualizing easily and fast without the need for sophisticated, time-consuming programming.
- MATLAB includes a large database of built-in algorithms and libraries. It also includes various embedded functions and tools; such as linear algebra function, neural network tools, probability functions, etc.
- The programmer can utilize the advanced programming techniques and object-oriented programming.
- The programmer can easily integrate MATLAB with other programming languages, or software. Also, the programmer is able to export to or import in any files between MATLAB and other software.

5. SIMULATION RESULTS

In this section, the results of gas allocation optimization using PSO algorithm are presented and discussed. Two different scenarios; a low-dimension problem with six wells, and a high-dimension problem with 56 wells are considered in our simulations. The constraints on the amount of available lift gas are considered (limited amount of lift gas is available). The optimization is implemented on the datasets from Buitrago et al. research. As mentioned, the ANN approach is employed to estimate the performance diagrams of the lift gas. The objective in the constrained optimization problem is to maximize oil production. The upper limit for the gas consumption is only considered as a constraint, and the gas consumption is not a term in the objective function. The objective function and the constraints equation is as (5).

$$maxZ = \sum_{i=1}^{\#ofwells} Q_{o_i}$$
$$\sum_{i=1}^{\#ofwells} Q_{g_i} \leq AG \qquad (5)$$

The simulation results for the six-well problem and 56-well problem, using the proposed approach and GA, are shown in Tables 1 and 2, respectively.

Moreover, the estimation of performance plots using the neural network approach are illustrated in Figure 1.

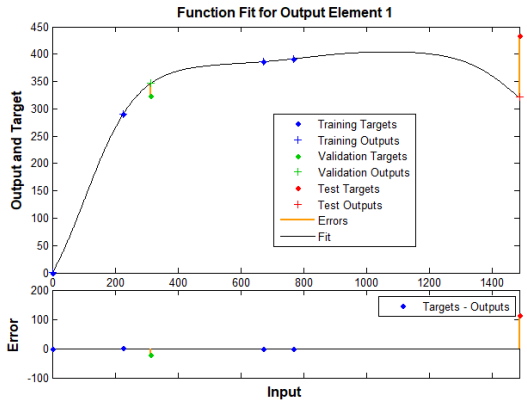


Figure 1: The performance plot estimates through the ANN algorithm.

Table 1: Simulation results on a 6-well problem using the propose method and GA

Well 1	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	296.6	335
GA	296.6	335
Well 2	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	507.7	720
GA	507.7	720
Well 3	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	931.7	1079
GA	934.8	1079
Well 4	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	353.9	534
GA	353.9	534
Well 5	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	910.0	757
GA	910.0	757
Well 6	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	3000.0	3425
GA	3000.0	3425

Table 2: Simulation results on a 56-well problem using the propose method and GA

	$q_g(MSCF/D)$	$q_0(B/D)$
PSO	22500	22541
GA	22500	22541

From Table 1, the optimum oil production is 3425 barrels in the constrained optimization problem, in both PSO and GA approaches. In a 6-well optimization problem, the results from the two evolutionary algorithms GA and PSO is almost the same, since it is a low-dimension problem. Obviously, in a higher dimension optimization problem with more computational complications, the performance of the evolutionary optimization methods will be recognizably different. Comparing the results of simulations in a 56-well problem proved that the proposed evolutionary algorithms performed more than 3% (more than 700 barrels) better than the classic approaches. Therefore, if the higher the dimension of the problem, the significantly better performance will be attained using the evolutionary optimization algorithms compared to the classical methods.

Although the results from GA and PSO approaches are the same, we recommend PSO for the gas allocation optimization problem. To prove the superiority of PSO over GA, we have shown the number of iterations needed to solve the same problem using the two algorithms (Figure 2). Thus, from the iteration graph, PSO converges a lot faster (13 times faster) than GA and requires less number of iterations for solving the same optimization problem. So, the operational costs for solving the problem using GA is significantly more than the cost associated with PSO. The parameter update processes in

PSO enhances the convergence pace in the algorithm. The main drawback of GA in this regard is that it does not update its parameters, and it does not include any tunable parameter in its process.

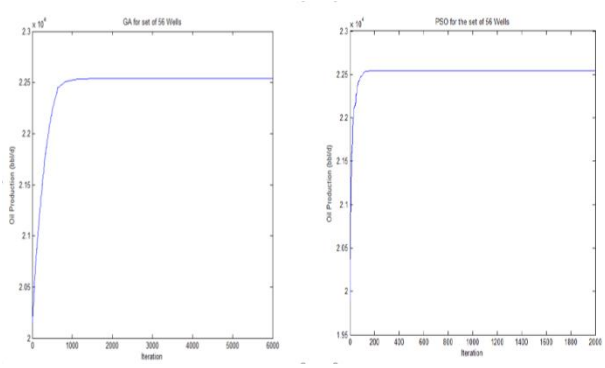


Figure 2: The number of iterations in PSO and GA for solving the 56-well problem

6. CONCLUSIONS

The gas distribution optimization problem is studied in this paper. The particle swarm optimization (PSO) approach is used for the first time for this problem. The performance plots are attained through an artificial neural network (ANN) learning. The proposed strategy is implemented on a high-dimensional (56-well) and a low-dimensional (6-well) problem. The better performance of the evolutionary optimization method (GA and PSO) over the classical approaches is more recognizable when the problem is of higher dimension (like the 56-well problem). PSO and GA showed similar performances; however, PSO performed much faster (13 times faster) and required less number of iterations than GA.

7. REFERENCES

[1] F. Rahmani, F. Razaghian, and A. Kashaninia, "High Power Two-Stage Class-AB/J Power Amplifier with High Gain and Efficiency," 2014.

[2] M. Ketabdar, "Numerical and Empirical Studies on the Hydraulic Conditions of 90 degree converged Bend with Intake," International Journal of Science and Engineering Applications, vol. 5, pp. 441-444, 2016.

[3] A. Hamedi, M. Ketabdar, M. Fesharaki, and A. Mansoori, "Nappe Flow Regime Energy Loss in Stepped Chutes Equipped with Reverse Inclined Steps: Experimental Development," Florida Civil Engineering Journal, vol. 2, pp. 28-37, 2016.

[4] R. Eini and A. R. Noei, "Identification of Singular Systems under Strong Equivalency," International Journal of Control Science and Engineering, vol. 3, pp. 73-80, 2013.

[5] Rostaghi-Chalaki, Mojtaba, A. Shayegani-Akmal, and H. Mohseni. "Harmonic analysis of leakage current of silicon rubber insulators in clean-fog and salt-fog." 18th International Symposium on High Voltage Engineering. 2013.

[6] Rahimikelarijani, Behnam, et al. "Optimal Ship Channel Closure Scheduling for a Bridge Construction." *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2017.

[7] F. Rahmani, F. Razaghian, and A. Kashaninia, "Novel Approach to Design of a Class-EJ Power Amplifier Using High Power Technology," World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, vol. 9, pp. 541-546, 2015.

[8] Rostaghi-Chalaki, Mojtaba, A. Shayegani-Akmal, and H. Mohseni. "A study on the relation between leakage current and specific creepage distance." 18th International Symposium on High Voltage Engineering (ISH 2013). 2013.

[9] M. Golan and C. H. Whitson, Well Performance, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, (1991) by Prentice-Hall. Inc.

[10] J. Redden, T. A. Sherman, and J. Blann, “Optimizing Gas-Lift Systems,” in Proceedings of Fall Meeting of the Society of Petroleum Engineers of AIME, 1974.

[11] T. D. Mayhill, “Simplified Method for Gas-Lift Well Problem identification and Diagnosis,” in Fall Meeting of the Society of Petroleum Engineers of AIME, 1974.

[12] E. Kanu, J. Mach, and K. Brown, “Economic Approach to Oil Production and Gas Allocation in Continuous Gas Lift (includes associated papers 10858 and 10865),” J. Pet. Technol., vol. 33, no. 10, pp. 1,887–1,892, Oct. 1981.

[13] N. Nishikiori, R. A. Redner, D. R. Doty, and Z. Schmidt, “An Improved Method for Gas Lift Allocation Optimization,” in Proceedings of SPE Annual Technical Conference and Exhibition, 1989.

[14] R. Eini, "Flexible Beam Robust Loop Shaping Controller Design Using Particle Swarm Optimization," Journal of Advances in Computer Research, vol. 5, pp. 55-67, 2014.

[15] R. Eini, and S. Abdelwahed. "Distributed Model Predictive Control Based on Goal Coordination for Multi-Zone Building Temperature." In *2019 IEEE Green Technologies Conference (GreenTech), Lafayette, LA*. 2019.

[16] B. T. Hyman, Z. Alisha, S. Gordon, "Secure Controls for Smart Cities; Applications in Intelligent Transportation Systems and Smart Buildings," International Journal of Science and Engineering Applications, vol. 8, pp. 167-171, 2019. doi: 10.7753/IJSEA0806.1004

[17] Heng, Li Jun, and Abesh Rahman. "Designing a robust controller for a missile autopilot based on Loop shaping approach." *arXiv preprint arXiv:1905.00958* (2019).

[18] Patel, Dev, Li Jun Heng, Abesh Rahman, and Deepika Bharti Singh. "Servo Actuating System Control Using Optimal Fuzzy Approach Based on Particle Swarm Optimization." *arXiv preprint arXiv:1809.04125* (2018).

[19] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, 1975.

[20] Bakker, V. Deljou, and J. Rahmani, "Optimal Placement of Capacitor Bank in Reorganized Distribution Networks Using Genetic Algorithm," International Journal of Computer Applications Technology and Research (IJCATR), vol. 8, pp. 2319-8656, 2019.

[21] F. Rahmani, "Electric Vehicle Charger based on DC/DC Converter Topology," International Journal of Engineering Science, vol. 18879, 2018.

[22] F. Rahmani and M. Barzegaran, "Dynamic wireless power charging of electric vehicles using optimal placement of transmitters," in 2016 IEEE Conference on Electromagnetic Field Computation (CEFC), 2016, pp. 1-1.

[23] M. Ketabdar and A. Hamedi, "Intake Angle Optimization in 90-degree Converged Bends in the Presence of Floating Wooden Debris: Experimental Development," Florida Civ. Eng. J, vol. 2, pp. 22-27.2016, 2016.

[24] M. Ketabdar, A. K. Moghaddam, S. A. Ahmadian, P. Hoseini, and M. Pishdadakhgari, "Experimental Survey of Energy Dissipation in Nappe Flow Regime in Stepped Spillway Equipped with Inclined Steps and Sill," International Journal of Research and Engineering, vol. 4, pp. 161-165, 2017

[25] A. Hamedi and M. Ketabdar, "Energy Loss Estimation and Flow Simulation in the skimming flow Regime of Stepped Spillways with Inclined Steps and End Sill: A Numerical Model," International Journal of Science and Engineering Applications, vol. 5, pp. 399-407, 2016.

[26] Rahimikelarijani, Behnam, Mohammad Saidi-Mehrabad, and Farnaz Barzinpour. "A mathematical model for multiple-load AGVs in Tandem layout." Journal of Optimization in Industrial Engineering (2019).

[27] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proceedings of ICNN’95 - International Conference on Neural Networks, 1995, vol. 4, pp. 1942–1948.

[28] N. Sfeir, H. Sharifi, "Internet of Things Solutions in Smart Cities," doi: 10.13140/RG.2.2.26015.51367 August 2019.

[29] H. Sharifi, "Singular Identification of a Constrained Rigid Robot," International Research Journal of Engineering and Technology (IRJET), vol. 5, pp. 941-946, 2018.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 358-362, 2019, ISSN:-2319–8656

Campus Placement Analyzer: Using Supervised Machine Learning Algorithms

Shubham Khandale
Student, M.Sc. (Big Data Analytics)
School of Computer Science, Faculty of Science
MIT-WPU, Pune, Maharashtra, India

Sachin Bhoite
Assistant Professor
School of Computer Science, Faculty of Science
MIT-WPU, Pune, Maharashtra, India

Abstract -- The main aim of every academia enthusiast is placement in a reputed MNC's and even the reputation and every year admission of Institute depends upon placement that it provides to their students. So, any system that will predict the placements of the students will be a positive impact on an institute and increase strength and decreases some workload of any institute's training and placement office (TPO). With the help of Machine Learning techniques, the knowledge can be extracted from past placed students and placement of upcoming students can be predicted. Data used for training is taken from the same institute for which the placement prediction is done. Suitable data pre-processing methods are applied along with the features selections. Some Domain expertise is used for pre-processing as well as for outliers that grab in the dataset. We have used various Machine Learning Algorithms like Logistic, SVM, KNN, Decision Tree, Random Forest and advance techniques like Bagging, Boosting and Voting Classifier and achieved 78% in XGBoost and 78% in AdaBoost Classifier.

Keywords: Pre-processing, Feature Selection, Domain expertise, Outliers, Bagging, Boosting, SVM, KNN, Logistics

1. INTRODUCTION

Nowadays Placement plays an important role in this world full of unemployment. Even the ranking and rating of institutes depend upon the amount of average package and amount of placement they are providing.

So basically main objective of this model is to predict whether the student might get placement or not. Different kinds of classifiers were applied i.e. Logistic Regression, SVM, Decision Tree, Random Forest, KNN, AdaBoost, Gradient Boosting and XGBoost. For this all over academics of students are taken under consideration. As placements activity take place in last year of academics so last year semesters are not taken under consideration

2. RELATED WORK

Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing, Feature.

Pothuganti Manvitha, Neelam Swaroopa (2019) used Random Forest and Decision Tree. The accuracy obtained after analysis for Decision tree is 84% and for the Random Forest is 86%. Hence, from the above-said analysis and prediction, it's better if the Random Forest algorithm is used to predict the placement results [1].

Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar(2017) used Decision Tree, Logistic Regression, Metabagging Classifier, Naïve Bayes and obtain highest 84.42% accuracy in Decision Tree. The objectives, which is to predict the placement status the students in Btech are most likely to have at the end of their final year placements. The accuracy of 71.66% with tested real-life data indicates that the system is reliable for carrying out its major objectives, which is to help teachers and placement cell[2].

Ajay Kumar Pal, Saurabh Pal (2013) they are predicting the placement of student after doing MCA by the three selected classification algorithms based on Weka. The best algorithm based on the placement data is Naïve Bayes Classification with an accuracy of 86.15% and the total time taken to build the model is at 0 seconds. Naïve Bayes classifier has the lowest average error at 0.28 compared to others.[3]

Syed A0068med, Aditya Zade, Shubham Gore, Prashant Gaikwad, Mangesh Kolhal (2017). Their objective is to analyze the previous year's student's historical data and predict placement chance of the current students and the percentage placement chance of the institution. They have used the Decision tree C4.5 Algorithm. Decision tree C4.5 algorithms are applied to the Company's previous year data & current requirement to generate the model and this model can be used to predict the students' eligibility in various companies. According to company eligibility criteria, they will send the notification to those candidates who are eligible for that campus interview and check the eligibility of candidate on the basis of percentage & technology [4].

Apoorva Rao r , Deeksha K C , Vishal Prajwal R , Vrushak K, Nandini M S (2018). They have used techniques like clustering along with that they have used classification rule Naïve Bayes algorithm that will classify students in five

different status i.e. Dream company, Core Company, Mass recruiters, Not eligible and Not interested[5]

3. DATASET DESCRIPTION AND SYSTEM FLOW

This approach was followed in following Figure 3.

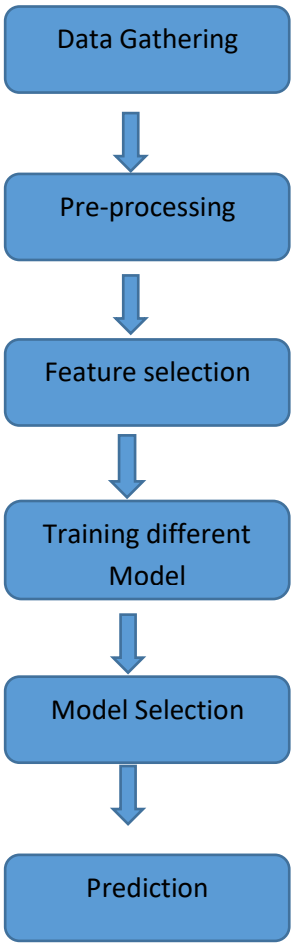


Figure 3. Flow chart

3.1 Data gathering and Pre-processing

The Data was collected from Training and placement department of MIT which consist of all the students of Bachler of Engineering (B.E) from 3 different colleges of their campus. The Data consists of 2338 records with 31 different attribute.

- Dataset contains academic information of students. As some students have completed their 12th and some of them are from diploma background who have directly taken admission to the second year so, we have merged 12th and diploma marks and made a single column for both.
- Some of the tuples where from M.tech background so we have dropped them and even in "current_aggregate" column we have dropped the NA values because the whole row was having NA.
- Replaced all NA values in columns "Current_Back_Papers", "Current_Pending_Back_Papers", all semester wise "Sem_Back_Papers", "Sem_Pending_Back_Papers" with 0 because it was null only if that student have no backlogs
- Using LabelEncoder from Preprocessing API in sklearn encoded the labels of columns "'Degree_Specializations", "Campus"," Gender", "year_down", "educational_gap"

3.2 Feature Selection

As per machine learning Feature Selection algorithms like "Ridge", "Lasso", "RFE", "plot importance", "F1 score" and "feature importance" we have got various outputs

- "Feature importance" with DT

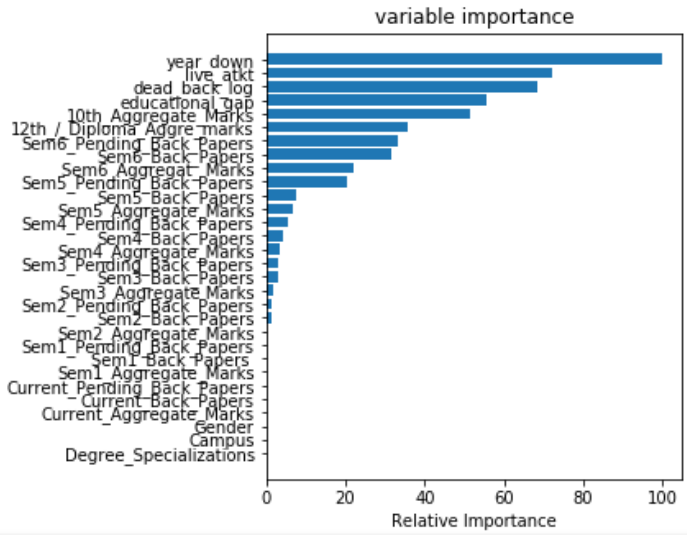


Figure 3.2.1 Feature importance with DT

- "Feature importance" with Random Forest

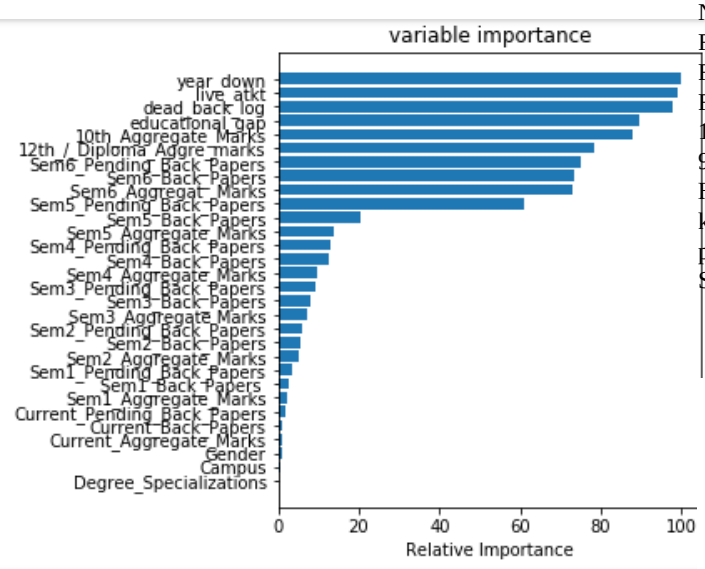


Figure 3.2.2 Feature importance with Random Forest

- "F1 score"

Feature Names	F1 score
Sem4_Aggregate_Marks	312.063809
Current_Aggregate_Marks	286.086537
Sem3_Aggregate Marks	255.771833
Sem2_Aggregate_Marks	164.183078
12th_/_Diploma_Aggre_marks	142.208129
Sem1_Aggregate_Marks	139.183936
Sem6_Aggregat_ Marks	136.333959
Sem5_Aggregate_Marks	131.988165
10th_Aggregate_Marks	128.526784
Sem6_Back_Papers	128.526784
live_atkt	47.908927
Sem5_Back_Papers	45.382049
Sem4_Back_Papers	43.547352

- **"RFE"**

Num Features: 5 Features support : [False False False False False False False False True False False True False False False True False True False False False False True False False False False False False] Features Ranking [25 6 4 3 24 8 13 22 1 10 11 1 23 17 2 1 19 1 5 18 7 21 1 26 16 20 15 12 14 9] selected Features:['Sem1_Pending_Back_Papers','Sem2_Pending_Back_Papers','Sem4_Aggregate_Marks','Sem4_Pending_Back_Papers', 'Sem6_Back_Papers']
Selected features index: [8, 11, 15, 17, 22]

- **"Ridge"**

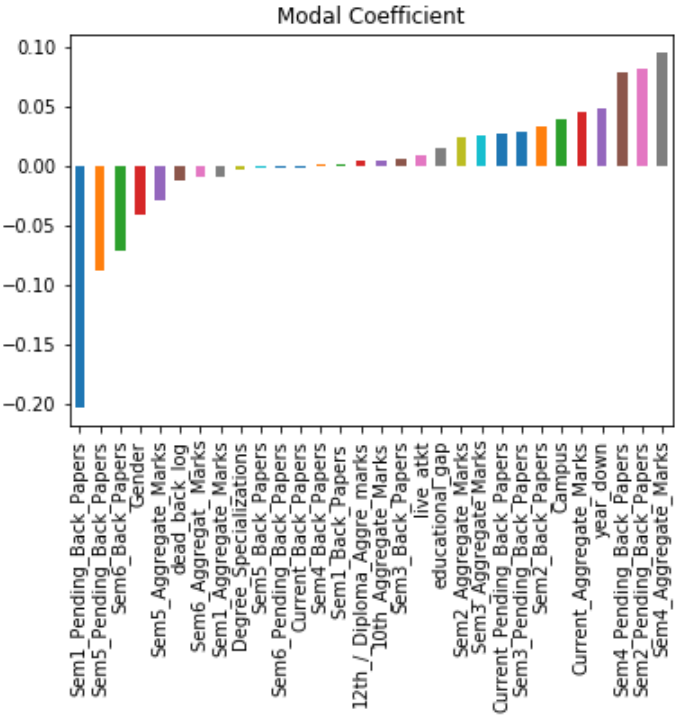


Figure 3.2.3 Feature Selection Using Ridge

- **"Lasso"**

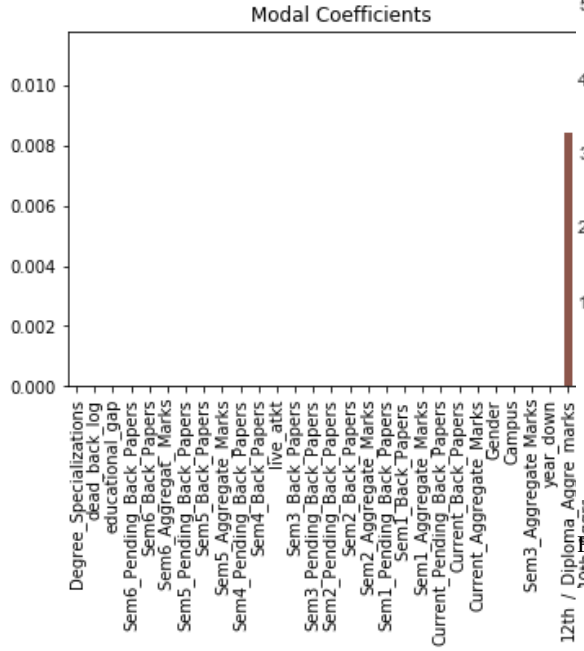


Figure 3.2.4 Feature Selection Using Lasso

But as per the domain knowledge we have selected all the features which are importance for our model

4. EXPLORATORY DATA ANALYSIS

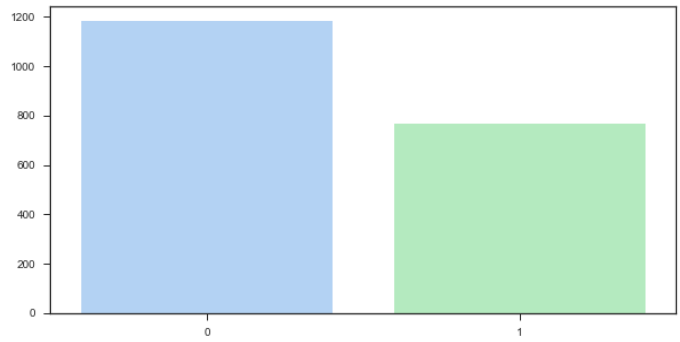


Figure 4.1 Total number of student placed

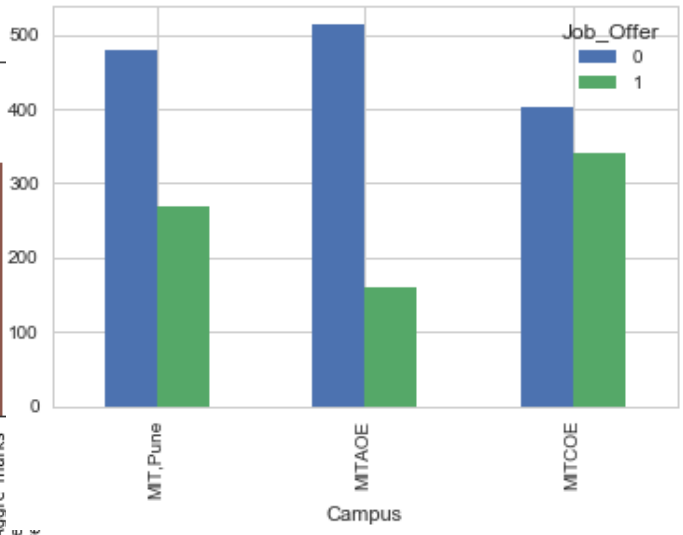


Figure 4.2 Campus wise number of students who got placed

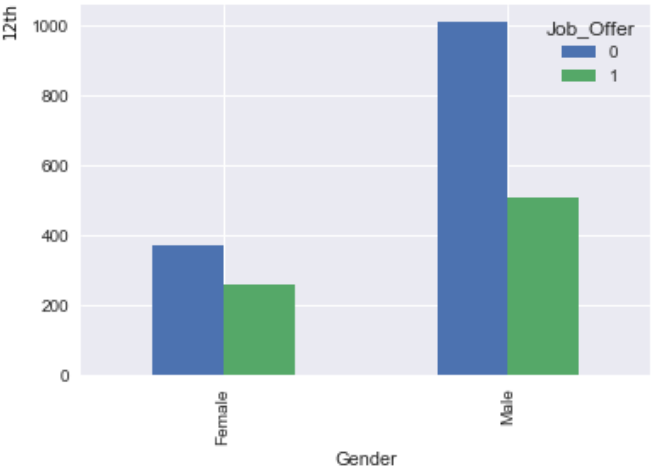


Figure 4.3 Gender Wise Student Placement

5. BAGGING AND BOOSTING

Bagging is nothing but bootstrap aggregating, it is an ensemble method to improve the accuracy and stability of the models. Random samples are taken with replacement and with every new sample that is generated is trained and the ensemble can make a prediction for the new instance by simply aggregating the prediction of all predictors

Boosting is nothing but the ensemble method that can combine different weak learner into a strong learner. Its main aim is to train predictors sequentially. Most popular are AdaBoost and Gradient Boosting.

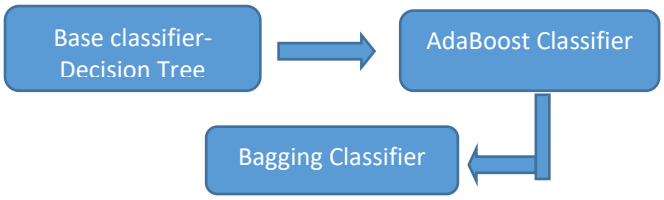


Figure 5.1 Layering of Classifiers

We have used Base Classifier as Decision Tree, over that we have used AdaBoost Classifier and over that we have used Baagging Classifier because we want to tune the accuracy of the model

6. RESULT AND CONCLUSION

Algorithms	**Accuracy**
Logistic Regression	58%
Support Vector Machine	69%
KNN	63.22 %
Decision Tree	69%
Random Forest	75.25%
AdaBoost(DT)	77%
Gradient Boosting	77%
Voting Classifier Soft	69.11%
Voting Classifier Hard	68.43%
XGBoost	78%

In this model, we have considered various academics records along with all semester's aggregate, live backlog, dead backlog, education gap, year down. This model will help the teachers to find whether the student will get placement or not prior in 3rd year only so that they can pay special attention to those students who are predicted as not getting placement. Even the institute can take major steps to improve the qualities of those students before their final placement. Various algorithms were used but the final model is selected on AdaBoost classifier along with the Bagging and Decision Tree as Base Classifier as its accuracy is very high.

The existing dataset was only for 3 colleges further even we can add more college's dataset to it for prediction. In future, we are going to implement Deep learning algorithms which may give better accuracy then Machine Learning models

7. REFERENCES

[1] Pothuganti Manvitha, Neelam Swaroopa "Campus Placement Prediction Using Supervised Machine Learning Techniques" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Sept 2019

[2] Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar "Student Placement Analyzer: A Recommendation System Using Machine Learning" 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2017), Coimbatore, INDIA, Jan. 06 – 07, 2017

[3] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of Students" I.J.Modern Education and Computer Science, 2013, 11, 49-56 Published Online, 11 November 2013

[4] Syed A0068med, Aditya Zade, Shubham Gore, Prashant Gaikwad, Mangesh Kolhal "Smart System for Placement Prediction using Data Mining" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653, Dec 2017

[5] Apoorva Rao r , Deeksha K C , Vishal Prajwal R , Vrushak K, Nandini M S "Student placement analyzer: a recommendation system using machine learning" ijariie-issn(o)-2395-4396, Jan 2018

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 363-366, 2019, ISSN:-2319–8656

Customer Churn Analysis and Prediction

Aditya Kulkarni [1]
M.sc (Big Data Analytics)
MIT WPU
Pune , India

Amruta Patil [2]
Msc (Big Data Analytics)
MIT WPU
Pune , India

Madhushree Patil [3]
Msc (Big Data Analytics)
MIT WPU
Pune , India

Sachin Bhoite [4]
Assistant Professor , Computer Science
MIT WPU
Pune , India

Abstract: When talking about any companies growth within market customers play an essential role in it , having the correct insights about customer behaviour and their requirements is the current need in this customer driven market . Preserving the interests of customers by providing new services & products helps in maintaining business relations . Customer churn is great problem faced by companies nowadays due to lagging in understanding their behaviour & finding solutions for it . In this project we have found causes of the churn for a telecom industry by taking into consideration their past records & then recommending them new services to retain the customers & also avoid churns in future . We used pie charts to check churning percentage later analysed whether there are ant outliers [using box plot] then dropped some features which were of less importance then converted all categorical data into numerical by using [Label Encoding for multiple category data & map function for two category data] plotted the ROC curve to get to know about true positive & false negative rate getting line at 0.8 then spitted the data using train test split .We used algorithms decision tree , Random Forest for feature selection wherein we got feature importance , then used logistic regression & found feature with highest weight assigned leading to cause of churn . Now in order to retain customers we can recommend them new services.

Keywords : Customer churn analysis telecom , Customer churn prediction & prevention , naïve bayes , logistic regression , decision tree , random forest

1.INTRODUCTION

The telecom industry is growing day by day hence user as well as operators are investing into this industry ,such a customer driven industry faces a huge financial issue if customer tend to leave their services . By using machine learning we can analyse , predict the way customer respond to these services , researches have proven that by using past data it could be accomplished [2] .

In this Customer Churn prediction & retention we are analysing the past behaviour of customers and accordingly finding the real cause of the churn , then predicting whether churn will happen in future by customers . By taking into account details like Monthly charges , services they have subscribed for , tenures , contract they will contribute into he end results i.e prediction.

Our aim is to use machine learning concepts to not only predict & retain customers but also to avoid further churns which would be beneficial to industry .

2.RELATED WORKS

We went through various articles & research papers , and then found that many researchers have worked on customer churn as it is a major problem faced by industries nowadays we found the following papers more promising

"A comparison of machine learning techniques for customer churn prediction Praveen Asthana has used decision tree , svm , naïve bayes , ANN & compared which model gives best accuracy and would help in prediction of customer churn to achieve better performance[1].

SCOTT A. NESLIN, SUNIL GUPTA, WAGNER KAMAKURA, JUNXIANG LU, and CHARLOTTE H. MASON* "Defection Detection: Measuring and

Understanding the Predictive Accuracy of Customer Churn Models" [2]here they have worked on measuring and increasing accuracy for churn prediction used logistic & tree approach .

We went through one more paper "Customer churn prediction in telecom using machine learning in big data platform" Abdelrahim Kasem Ahmad* , Assef Jafar and Kadan Aljoumaa [3] they have used decision tree , random forest , XGBoosting , they used this algorithm for classification in predictive churn of customers getting better accuracy.

S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." , here they have used predictive model in a bank with personalized action to retain customer & have also used recomender system[4] .

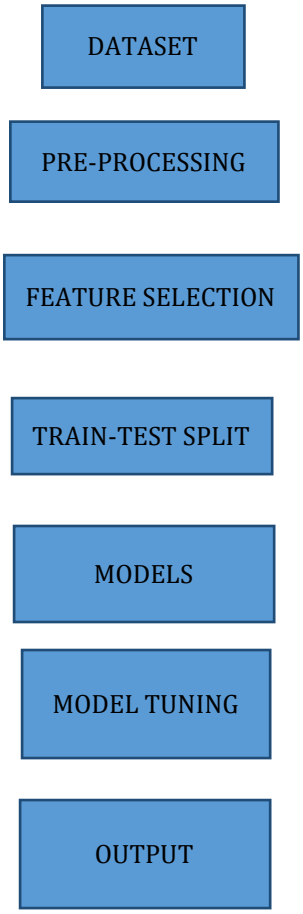
K. Coussement, and D. Van den Poel "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers.", they have used logistic regression , svm & random forest classification algorithms to filter out the churners from non-churner[5].

L Miguel APM. "Measuring the impact of data mining on churn management" , they have proposed a analysis framework which prefigure impact of data mining for churn management[6] .

Adnan Amin , Babar shah , Awais Adnan "Customer churn prediction in telecommunication industry using data certainty"[7], The dataset is grouped into different zones based on the distance factor which are then divided into two categories as data with high certainty, and data with low certainty, for predicting customers exhibiting Churn and Non-churn behaviour.

3. PROCESS FLOW

The data we got was mostly balanced & categorical data then we began with Data Cleaning, Pre-processing, removing unwanted columns, feature selection, label encoding.



3.1 DATASET

We took this telecom dataset from online website source took all the insights regarding the data .
Attributes of the dataset : :
Customerid, gender, SeniorCitizen, Partner, Dependents,tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn.

3.2 DATA PRE-PROCESSING

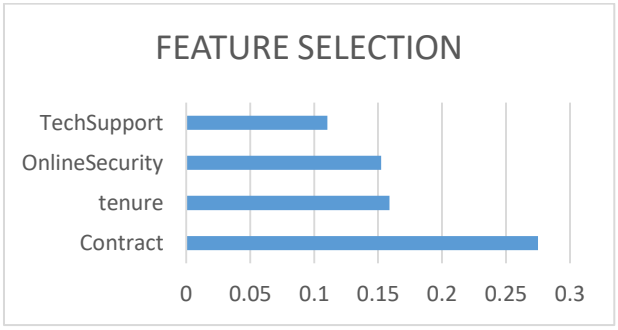
Data pre-processing is important task in machine learning. It converts raw data into clean data. Following are technique, we have applied on data: -

- Missing Values – Here we had missing values in Totalcharges feature which we then eliminated and adjusted them with mean values . These are the missing row values within data if not handled would later lead to errors for converting data type as it takes string value for empty spaces .
- Label Encoder – For categorical variables this is perfect method to convert them into numeric values , best used when having multiple categories . We had various categorical values converted them into numeric for further use in algorithms .
- Drop Columns – As we took insights from the data we came to know some of the features were of less importance so we dropped them to reduce number of features .

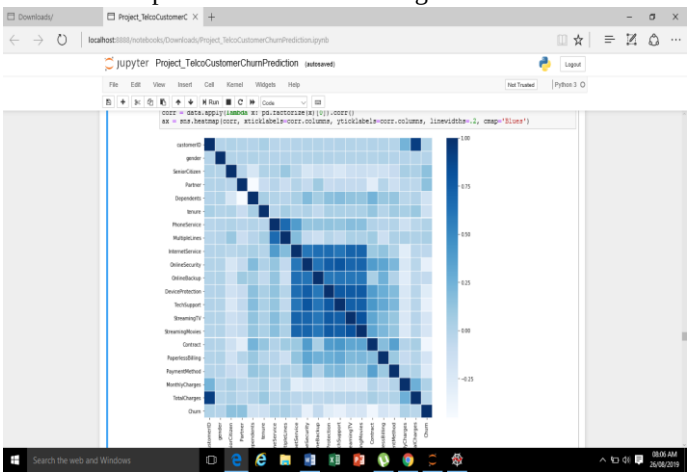
3.3 FEATURE SELECTION

As we had number of features and most of them were of great importance so we used feature section to get to know which of them are contributing towards the accuracy of the model .

We used Decision tree , Random forest for feature selection so using decision tree we got accuracy[80] and by using arandom forest we go t [80%] so random forest gave us four features
Index(['tenure', 'Contract', 'MonthlyCharges', 'TotalCharges'], dtype='object')
[(0.2251735641431145, 'Contract'), (0.1687558104226648, 'tenure'), (0.12539865168020692, 'OnlineSecurity'), (0.1128092761196452, 'TechSupport'), (0.10731999001345587, 'TotalCharges'), (0.08573112448285626, 'MonthlyCharges'),
Here we can see contarct is having more importance resulting factor for churn .

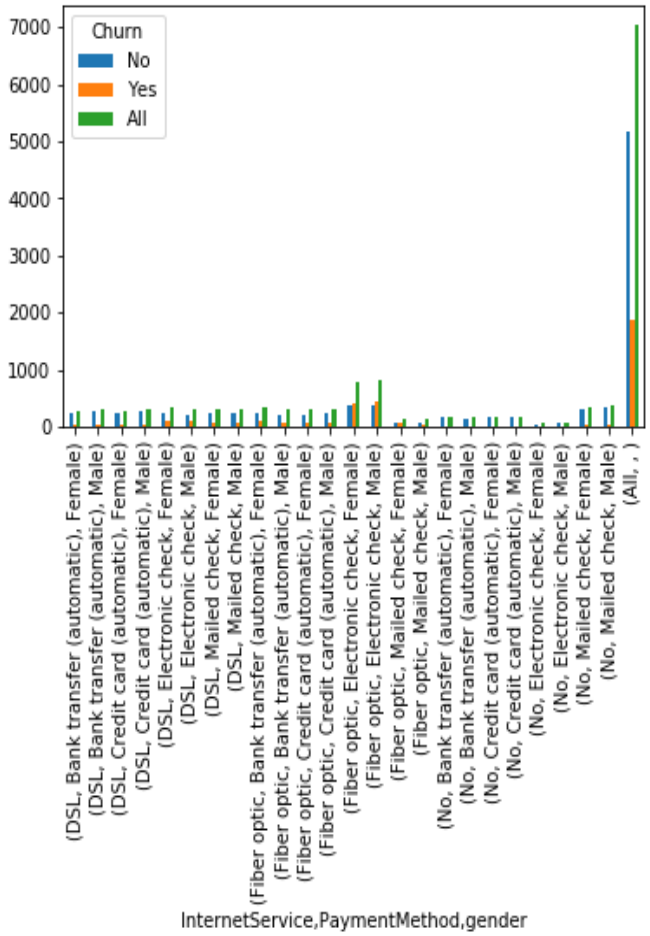


And for decision tree we got accuracy of [77%]
Used heat Map for correlation checking :

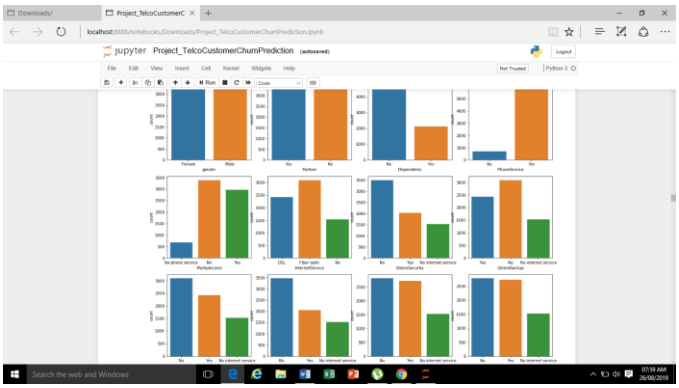


4. EXPLORATORY DATA ANALYSIS

In this phase we will look towards those features which we didn't consider in feature selection but are contributing factor for prediction .

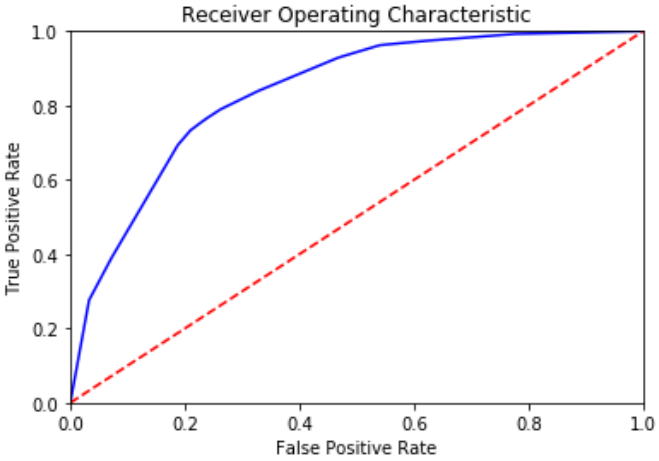


Here we can see that customer who took fibre optics for month-to-month contract whether it be male/female resulted in churn.



Also visualised all the features within the dataset & came to know the distributions .

Got roc curve ;



Confusion Matrix :
[[3816 295]
[924 590]] .

5. RESULT AND DISCUSSION

Now after all the cleaning up & pre-processing of the data now we separate our data for further applying algorithms on it. By using :

1. Train-Test Split
2. Modeling
3. Tuning Model

5.1 Train-Test Split:

To create the model we train our dataset while testing data set is used to test the performance. So, in our data, we have split into 80% for training data and 20% for testing data because it makes the classification model better whilst more test data makes the error estimate more accurate.

5.2 Modelling :

Following are model, we applied to check which model gives better accuracy:

- Support Vector Classifier (SVC):
  This algorithm is used for classification problem. The main objective of SVC is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, when obtaining the hyperplane, you'll then feed some options to your category to examine what the "predicted" class is.

- Decision Tree:
  Decision tree is non-parametric supervised learning. It is used for both classification and regression problem. It is flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path between root and leaf represent classification rules. It creates a comprehensive analysis along with each branch and identifies decision nodes that need further analysis.

- Random Forest:
  Random Forest is a meta estimator that uses the number of decision tree to fit the various sub samples drawn from the original dataset. we also can draw data with replacement as per the requirements.

- K-Nearest Neighbours (KNN):

K-Nearest Neighbours (KNN) is supervised learning algorithm which is used to solve regression and classification problem both. Where 'K' is number of nearest neighbours. It is simple to implement, easy to understand and it is lazy algorithm. Lazy algorithm means it does not need any training data points for model generation . All training data used in the testing phase.

- Naïve Bayes:
  A Naive Bayes Classifier is a supervised machine learning algorithm which uses the Bayes' Theorem, that features are statistically independent. It finds many uses in the probability theory and statistics. By simple machine learning problem, where we need to learn our model from a given set of attributes (in training examples) and then form a hypothesis or a relation to a response variable.

  - Logistic Regression :

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes . Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Models used & their accuracy : :

Model	Accuracy
Logistic Regression	80.38%
Decision Tree	77.81%
Random Forest Tree	80.02%
Naïve Bayes	74.91%
SVM	80.1%
K – Nearest Neighbour	76.61%
XGBoost	80%

Figure 5: Accuracy for Different Models

5.3 MODEL TUNING:

Here we tune the model to increase model performance without overfitting the model.

- XGBoost **:**
  XGBoost stands for extreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance[3].

We used XGBoost to check the error function and reduce it [Accuracy :80%] , by using cross validation checked for the reducing RMSE also it handles the missing values , initially our RMSE was [[0]validation_0-error: 0.208955] later it came [[10] validation_0-error: 0.200426]

6. CONCLUSION

Here we had past records of customers who had churned and using that data we predicted whether new customer would tend to churn or not , this will help the companies to get to know the behaviour of customer & how to maintain their interests into the services of company . Further the company can also use recommender system to retain customers and also avoid the further churns . We used various algorithms wherein Logistic regression gave us high accuracy close to this accuracy were Random Forest , SVM .

The dataset did not consisted of records which would tell us whether customer has switched the services , that will help in recommending new services further . Now we are going to build a recommender system to avoid churns & retain the old customers .

7. REFERENCES

[1] Praveen Ashtana "A comparison of machine learning techniques for customer churn prediction" International Journal of Pure and Applied Mathematics Volume 119 No. 10 2018, 1149-1169 ISSN: 1311-8080

[2] SCOTT A. NESLIN, SUNIL GUPTA, WAGNER KAMAKURA, JUNXIANG LU, and CHARLOTTE H. MASON* "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models" Journal of Marketing Research 204 Vol. XLIII (May 2006), 204–211 , ISSN: 0022-2437.

[3] Abdelrahim Kasem Ahmad* , Assef Jafar and Kadan Aljoumaa "Customer churn prediction in telecom using machine learning in big data platform" - Journal of Big Data **volume 6**, Article number: 28 (2019) , published on 20th March 2019 .

[4] S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." Expert Systems with Applications, vol. 31, no. 3, pp. 515–524, 2006.

[5] K. Coussement, and D. Van den Poel. "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers." Expert Systems with Applications, vol. 36, no. 3, pp. 6127–6134, 2009

[6] L. Miguel APM. "Measuring the impact of data mining on churn management." Internet Research, vol. 11, no. 5, pp. 375–387,2001

[7] Amin , Babar shah , Awais Adnan "Customer churn prediction in telecommunication industry using data certainty" Journal of business research Volume 94, January 2019, Pages 290-301.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656

Air Quality Prediction using Machine Learning Algorithms

Pooja Bhalgat
Student
M.Sc(Big Data Analytics)
MIT-WPU, Pune, India

Sejal Pitale
Student
M.Sc(Big Data Analytics)
MIT-WPU, Pune, India

Sachin Bhoite
Assistant Professor
Computer Science
MIT-WPU, Pune, India

Abstract: :Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution.With this increasing air pollution,Weare in need of implementing models which will record information about concentrations of air pollutants(so2,no2,etc).The deposition of this harmful gases in the air is affecting the quality of people's lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available.In this paper, machine learning techniques are used to predict the concentration of so2 in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs.Models in time series are employed to predict the so2 readings in nearing years or months.

Keywords: Machine Learning, Time Series, Prediction, Air Quality, SO2

1. INTRODUCTION

In the developing countries like India, the rapid increase in population and economic upswing in cities have lead to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has direct impact on humans health .There has been increased public awareness about the same in our country.Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution. Precised air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society.

Sulphur Dioxide is a gas. It is one of the major pollutants present in air.It is colorless and has a nasty, sharp smell.It combines easily with other chemicals to form harmful substances like sulphuric acid, sulfurous acid etc. Sulfur dioxide affects human health when it is breathed in. It irritates the nose, throat, and airways to cause **coughing, wheezing, shortness of breath**, or a tight feeling around the chest. The concentration of sulphur dioxide in the atmosphere can influence the **habitat suitability** for plant communities, as well as animal life.

The proposed system is capable of predicting concentration of Sulphur Dioxide for forthcoming months / years.

2. RELATED WORK

In this research paper the students have forecasted the air quality of India by using machine learning algorithms to predict the air quality index(AQI) of a given area. Air quality Index is a standard measure to determine the quality of air. Concentrartion of Gases such asso2, no2,co2, rspm, spm. etc. are recorded by the agencies . These students have developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. They improved the efficiency of the model by applying cost Estimation for predictive Problem. They say that this model is capable of successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration.[1]

This paper presents an integrated model using Artificial Neural Networks and Kriging to predict the level of air pollutants at various locations in Mumbai and Navi Mumbai using past data available from meteorological department and Pollution Control Board. The proposed model is implemented and tested using MATLAB for ANN an R for Kriging and the results are presented.[2]

This system has used the Linear regression and Multilayer Perceptron (ANN)Protocol for prediction of the pollution of next day. The system helps to predict next date pollution details based on basic parameters and analyzing pollution details and forecast future pollution. Time Series Analysis was also used for recognition of future data points and air pollution prediction.[3]

This proposed system does two important tasks (i). Detects the levels of PM2.5 based on given atmospheric values. (ii) Predicts the level of PM2.5 for a particular date. Logistic regression is used to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. The primary goal is to predict air pollution level in City with the ground data set.[4]

The major objective of this paper was to provide a snapshot of the vast research work and useful review on the current state-of-the-art on applicable big data approaches and machine learning techniques for air quality evaluation and predication.Air quality map were illustrated and visualized using data from Shenzhen, China. Artificial neural network

(ANN), Genetic Algorithm ANN Model, Random forest, decision tree, Deep belief network are the algorithms which were used and various pros and cons of the model were presented.[5]

.

3. DATASET

3.1 Dataset/Source: Kaggle
Structured/Unstructured data: Structured Data in CSV format.

Dataset Description:
The dataset consists of around 450000 records of all the states of India.We worked only on Dataset of Maharashtra.So we had 60383 records. This dataset consist of 13 attributes listed below.

1)stn_code
2)sampling_date
3) state
4) location
5) agency
6)type
7)so2
8)no2
9)rspm
10) spm
11)location_monitoring_station
12)pm2_5
13)date

Station code is a code given to each station that recorded the data,sampling date is the date when the data is recorded.state and location represents state and cities whose data is recorded and agency is the name of agency that recorded the data.Type states the type of area where the data was recorded such as industrial,residential,etc.so2,no2,rspm and spm is the amount of sulphur dioxide, nitrogen dioxide, respirable suspended particulate matter and suspended particulate matter measured respectively.date is a cleaner version of sampling_date. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.But majority of values in this column are null.

Splitting for Testing : Data Splitting was done as 80% for training and 20% for testing.

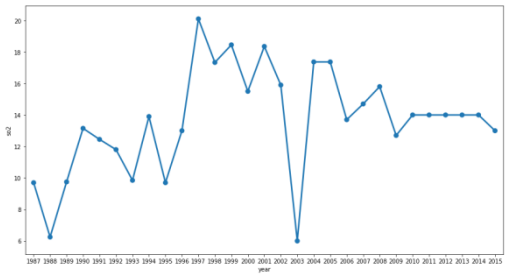
Preprocessing and Feature Selection:
We only studied and applied algorithms on the data of Maharashtra State .Hence, no. of rows was reduced to 60,383 and state column automatically is of no more use.
All the values in pm2_5 were null values ,so we dropped the column.The agency's name have nothing to do with how much polluted the state is. Similarly, stn_code is also not useful.
The date is a cleaner representation of sampling_date attribute and so we will eliminate the redundancy by removing the latter. location_monitoring_station attribute is again unnecessary as it contains the location of the monitoring station which we do not need to consider for the analysis.

So, to summarize we have deleted the following features from our dataset :

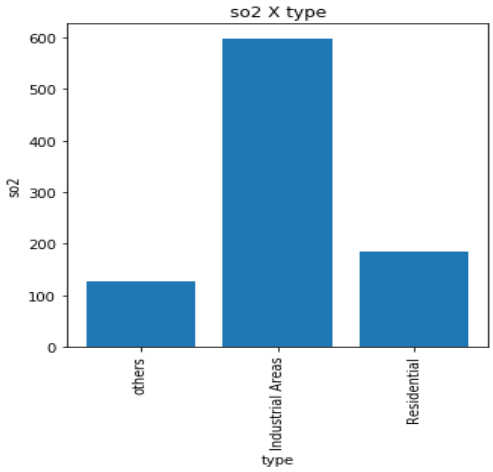
state,pm2_5,agency, stn_code, sampling_date and location_monitoring_station
We have simplified the type attribute to contain only one of the three categories: industrial, residential, other.For SO2 and NO2, we replaced nan values by mean.For date, we have dropped nan values as there were only 3 null values.
So after pre-processing our dataset contains 60,380 rows and 7 columns.

4. EXPLORATORY DATA ANALYSIS:

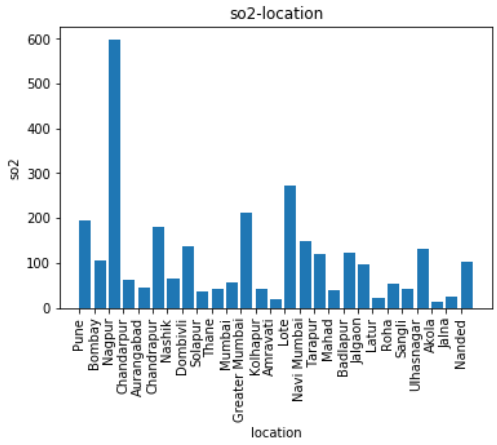
- The below graph shows concentration of so2 over the years.It was highest in the years of 1997 and 2001 and lowest in the years 1988 and 2003 .However,it is stable for the latest years.



- This graph shows that the amount of so2 is highest in the industrial areas.



- From this graph we can conclude that Nagpur has the deadliest amount of so2 as compared to other cities whereas Akole , Amravati are sparsely polluted followed by Jalna and Kolhapur.



5. RESULT AND DISCUSSION:

We are able to identify the future data points using Time Series Analysis.
Models used for the same are :

1)AR model:(autoregressive model)

Test MSE: 166.358
Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.
It is a very simple idea that can result in accurate forecasts on a range of time series problems.

yhat = b0 + b1*X1

Where yhat is the prediction, b0 and b1 are coefficients found by optimizing the model on training data, and X is an input value.
This technique can be used on time series where input variables are taken as observations at previous time steps, called lag variables.
For example, we can predict the value for the next time step (t+1) given the observations at the last two time steps (t-1 and t-2). As a regression model, this would look as follows:

X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)
Because the regression model uses data from the same input variable at previous time steps, it is referred to as an autoregression (regression of self).[6]

2)ARIMA MODEL:

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.

ARIMA is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:
p: The number of lag observations included in the model, also called the lag order.
d: The number of times that the raw observations are differenced, also called the degree of differencing.
q: The size of the moving average window, also called the order of moving average.[7]

6. CONCLUSION

Based on the bar plots plotted we come to the conclusion that some cities are highly polluted and need urgent attention. Also for cities like Pune ,Mumbai where concentration of so2 is increasing, we can take measures from now to not face problems later.We used AR model and ARIMA model for predicting values of so2. Features such as location_monitoring_station or station code were of no use as they have nothing to do with so2 predictions.

So2 safe levels are as follows:
0.20 ppm (parts per million) averaged over a one hour period. 0.08 ppm averaged over a 24 hour period. 0.02 ppm averaged over a one year period.

In order to predict air quality, pm2_5 is also an important attribute. The values of this must be recorded in future as this particulates are responsible for various health effects including cardiovascular effects such as cardiac arrhythmias and heart attacks, and respiratoryeffects such as asthma attacks and bronchitis.

This model is not able to show expected output as the data is not in sequence as per date column.The same is the problem for cities.If we predict for the entire state, it wont be helpful So we will be now calculating AQI and use classification models further.

This model further, also makes us aware of the challenges in future and research needs such as pm2.5,AQI,etc.

7. REFERENCES

[1] Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. “Indian Air Quality Prediction And Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue)
[2] Suhasini V. Kottur , Dr. S. S. Mantha. “An Integrated Model Using Artificial Neural Network

(Ann) And Kriging For Forecasting Air Pollutants Using Meteorological Data". International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940 Vol. 4, Issue 1, January 2015

[3] RuchiRaturi, Dr. J.R. Prasad ."Recognition Of Future Air Quality Index Using Artificial Neural Network".International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[4] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu ." Detection and Prediction of Air Pollution using Machine Learning Models". International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

[5] Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie." Air Quality Prediction: Big Data and Machine Learning Approaches". International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

[6] https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

[7] https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 371-374, 2019, ISSN:-2319–8656

Individual Household Electric Power Consumption Forecasting using Machine Learning Algorithms

Aaditi Parate
Student, M.Sc. (Big Data Analytics)
MIT WPU Pune, India

Sachin Bhoite
Assistant Professor, Computer Science
MIT-WPU, Pune, India

----------**********----------

Abstract: Electric energy consumption is the actual energy demand made on existing electricity supply. However, the mismanagement of its utilisation can lead to a fall in the supply of electricity. It is therefore imperative that everybody should be concerned about the efficient use of energy in order to reduce consumption [1]. The purposes of this research are to find a model to forecast the electricity consumption in a household and to find the most suitable forecasting period whether it should be in daily, weekly, monthly, or quarterly. The time series data in our study is the individual household electric power consumption [4].To explore and understand the dataset I used line plots for series data and histograms for the data distribution. The data analysis has been performed with the ARIMA (Autoregressive Integrated Moving Average) model.

Keywords: Energy consumption prediction, ARIMA, AR, MA, Python.

---

1. INTRODUCTION

Electricity load forecasting has gained substantial importance nowadays in the modern electrical power management systems with elements of smart greed technology. A reliable forecast of electrical power consumption represents a starting point in policy development and improvement of energy production and distribution. At the level of individual households, the ability to accurately predict consumption of electricity power significantly reduces prices by appropriate systems for energy storage. Therefore, the energy efficient power networks of the future will require entirely new ways of forecasting demand on the scale of individual households [2]. The analysis of a time series used forecasting techniques to identify models from the past data. With the assumption that the information will resemble itself in the future, we can thus forecast future events from the occurred data. There are several techniques of forecasting and these techniques provide forecasting models of different accuracy. The accuracy of the prediction is based on the minimum error of the forecast. The appropriate prediction methods are considered from several factors such as prediction interval, prediction period, characteristic of time series, and size of time series [4].

In this research, we are interest in time series analysis with the popular forecasting technique that I used in this study; ARIMA (Autoregressive Moving Average) I applied this method for detecting patterns and trends of the electric power consumption in the household with real time series period in daily, weekly, monthly, and quarterly. I used Python program for constructing the model.

2. RELATED WORK:

How to Load and Explore Household Electricity Usage Data In this tutorial, you will discover a household power consumption dataset for multi-step time series forecasting and how to better understand the raw data using exploratory analysis.[5]

How to Develop an Autoregression Forecast Model for Household Electricity Consumption In this tutorial, you will discover how to develop and evaluate an autoregression model for multi-step forecasting household power consumption.[6]

Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models: In this research, we are interest in time series analysis with the most popular method, that is, the Box and Jenkins method. The result model of this method is quite accurate compared to other methods and can be applied to all types of data movement. There were two forecasting techniques that were used in this study; Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Moving Average (ARMA).[1]

3. DATASET DESCRIPTION:

Source:
https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption

The Household Power Consumption dataset is a multivariate time series dataset that describes the electricity consumption for a single household over four years.

This archive contains 2075259 measurements gathered in a house located in Sceaux (7km of Paris, France) between December 2006 and November 2010 (47 months).

It is a multivariate series comprised of seven variables (besides the date and time); they are:

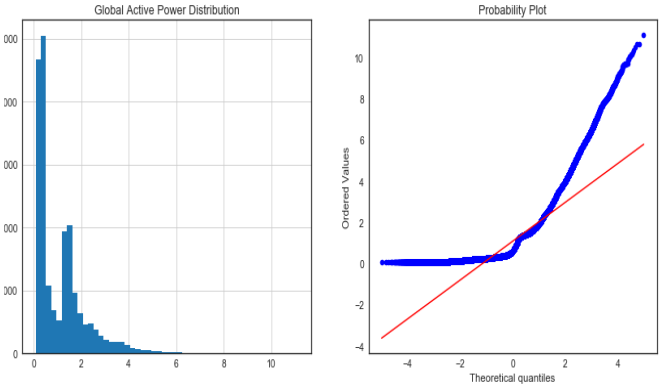
- **global_active_power**: The total active power consumed by the household (kilowatts).
- **global_reactive_power**: The total reactive power consumed by the household (kilowatts).
- **voltage**: Average voltage (volts).
- **global_intensity**: Average current intensity (amps).
- **sub_metering_1**: Active energy for kitchen (watt-hours of active energy).
- **sub_metering_2**: Active energy for laundry (watt-hours of active energy).
- **sub_metering_3**: Active energy for climate control systems (watt-hours of active energy).

4. PRE-PROCESSING:

The dataset contains some missing values in the measurements (nearly 1,25% of the rows). All calendar timestamps are present in the dataset but for some timestamps, the measurement values are missing: a missing value is represented by the absence of value between two consecutive semi-colon attribute separators. For instance, the dataset shows missing values on April 28, 2007.We cannot ignore the missing values in this dataset therefore we cannot delete the missing values. I copied the observation from the same time the day before and implemented this in a function named *fill_missing()* that will take the NumPy array of the data and copy values from exactly 24 hours ago Then we saved cleaned-up version of the dataset to a new file *household_power_consumption.csv*‘[3].

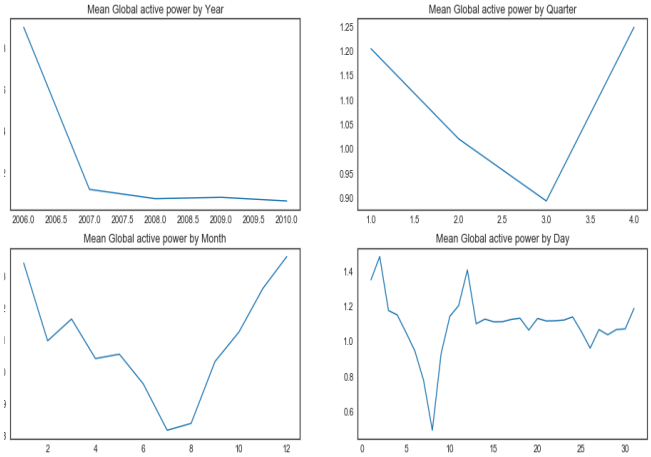
5. EXPLORATORY DATA ANALYSIS:

1) Global active power distribution plots:



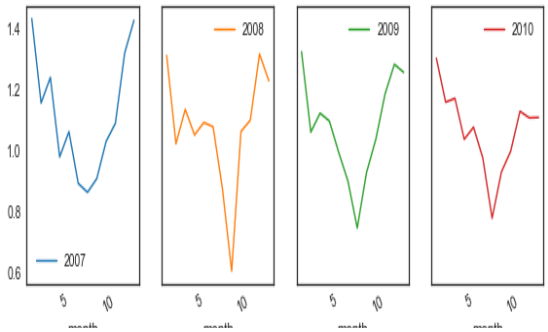
Normal probability plot also shows the data is far from normally distributed.

2) 1st graph represents the mean global active power by Year.
2nd graph represents the mean global active power by Quarter.
3rd graph represents the mean active power by the Month and the 4th graph represents the mean global active power by Day.



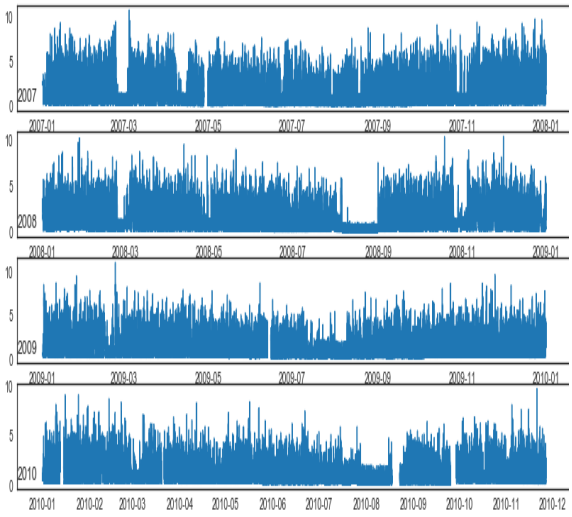
The above plots confirmed our previous discoveries. By year, it was steady. By quarter, the lowest average power consumption was in the 3rd quarter. By month, the lowest average power consumption was in July and August. By day, the lowest average power consumption was around 8th of the month

3) Global active power by year. This time we removed year 2006

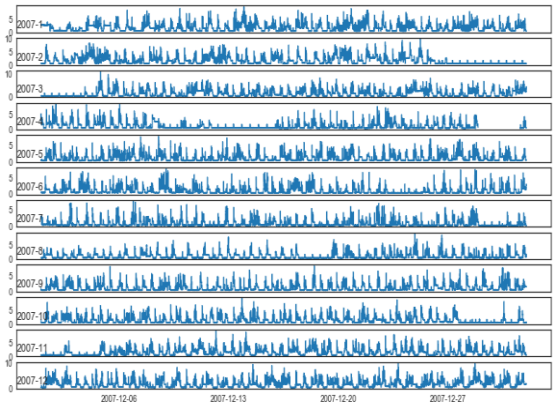


The pattern is similar every year from 2007 to 2010.

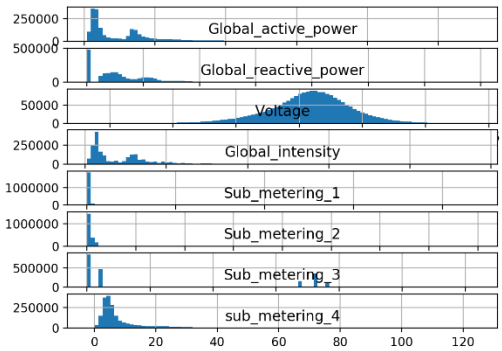
4) Line plot of Active power for years:



5) Line plots for Active Power for all months in a year:



6) Histogram plots for Each Variable in the Power Consumption Dataset



[4].

5. RESULT AND DISCUSSION:

I developed an autoregression model for univariate series of daily power consumption. I used the Statsmodels library that provides multiple ways of developing an AR model, such as using the AR, ARMA, ARIMA, and SARIMAX classes.

I use the ARIMA implementation as it allows for easy expandability into differencing and moving average.

First, the history data comprised of weeks of prior observations is converted into a univariate time series of daily power consumption. I specified an AR (7) model, which in ARIMA notation is ARIMA(7,0,0).

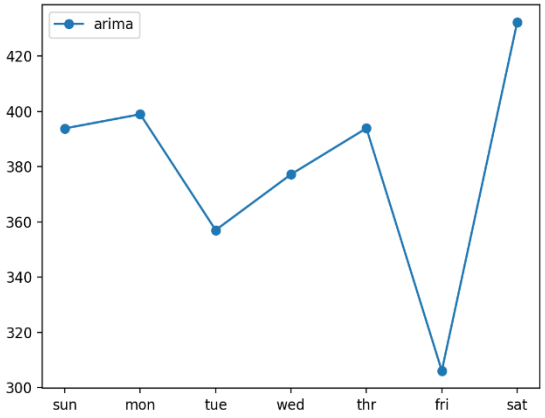
Running the example first prints the performance of the AR (7) model on the test dataset.

We can see that the model achieves the overall RMSE of about 381 kilowatts.

This model has skill when compared to naive forecast models, such as a model that forecasts the week ahead using observations from the same time one year ago that achieved an overall RMSE of about 465 kilowatts.

A line plot of the forecast is also created, showing the RMSE in kilowatts for each of the seven lead times of the forecast. We can see an interesting pattern. We might expect that earlier lead times are easier to forecast than later lead times, as the error at each successive lead time compounds.

Instead, we see that Friday (lead time +6) is the easiest to forecast and Saturday (lead time +7) is the most challenging to forecast. We can also see that the remaining lead times all have a similar error in the mid- to high-300 kilowatt range. [3]



6. CONCLUSION:

Many researchers wrote about the ARIMA model, AR model, MA model and also worked with these models on the consumption of electricity. I find the ARIMA model easy as compared to the other models and also the ARIMA model gives a better accuracy then the other models.

Therefore in this research I used the ARIMA model on the individual household electricity consumption dataset. Then, chose the suitable forecasting method and identified the most suitable forecasting period by considering the smallest values of RMSE. In this data set the consumption of electricity is more in the month of December and regular during the other time period of the year.The results showed that the ARIMA model represent the most suitable forecasting periods in monthly and quarterly, daily and weekly.

The ARIMA model result: arima: [465.902] 428.0, 448.9, 395.8, 522.3, 450.4, 380.5, 597.9Here the RMSE is 465.902

7. REFERENCES:

[1] c Kamunda A Study on Efficient Energy Use for Household Appliances in Malaw

[2] Naser Farag Abed1 and Milan M.Milosavljevic 1,2 1 Singidunum University Belgrade, 11000, Serbia 2 School of Electrical Engineering, Belgrade University, Belgrade, 11000, Serbia“Single Home Electricity Power Consumption Forecast Using Neural Networks Model”IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 1, January 2016. ISSN 2348 – 7968

[3]https://machinelearningmastery.com/how-to-develop-an-autoregression-forecast-model-for-household-electricity-consumption/#

[4] Pasapitch Chujai*, Nittaya Kerdprasop, and Kittisak Kerdprasop“Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models”Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

[5] https://machinelearningmastery.com/how-to-load-and-explore-household-electricity-usage-data/

[6] https://machinelearningmastery.com/how-to-develop-an-autoregression-forecast-model-for-household-electricity-consumption/#:

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 375-378, 2019, ISSN:-2319–8656

Restaurants Rating Prediction using Machine Learning Algorithms

Atharva Kulkarni[1]
Student, M.Sc (BDA)
MIT-WPU
Pune, Maharashtra, India

Divya Bhandari[2]
Student, M.Sc (BDA)
MIT-WPU
Pune, Maharashtra, India

Sachin Bhoite[3]
Assistant Professor
Department of Computer Science, MIT-WPU
Pune, Maharashtra, India

Abstract: Restaurant Rating has become the most commonly used parameter for judging a restaurant for any individual. A lot of research has been done on different restaurants and the quality of food it serves. Rating of a restaurant depends on factors like reviews, area situated, average cost for two people, votes, cuisines and the type of restaurant.
The main goal of this is to get insights on restaurants which people like visit and to identify the rating of the restaurant. With this article we study different predictive models like Support Vector Machine (SVM),Random forest and Linear Regression, XGBoost, Decision Tree and have achieved a score of 83% with ADA Boost.

Key Words: Pre-processing, EDA, SVM Regressor, Linear Regression, XGBoost Regressor, Boosting.

1. INTRODUCTION

Zomato is the most reputed company in the field of food reviews. Founded in 2008, this company started in India and now is in 24 different countries. Its is so big that the people now use it as a verb. “Did you know about this restaurant? Zomato it”. The rating is the most important feature of any restaurant as it is the first parameter that people look into while searching for a place to eat. It portrays the quality, hygiene and the environment of the place. Higher ratings lead to higher profit margins. Notations of the ratings usually are stars or numbers scaling between 1 and 5.

Zomato has changed the way people browse through restaurants. It has helped customers find good places with respect to their dining budget.

Different machine learning algorithms like SVM, Linear regression, Decision Tree, Random Forest can be used to predict the ratings of the restaurants.

2. RELATED WORK

Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing and feature selection.

[1] Shina, Sharma S. and Singha A. have used Random forest and decision tree to classifying restaurants into several classes based on their service parameters. Their results say that the Decision Tree Classifier is more effective with 63.5% of accuracy than Random Forest whose accuracy is merely 56%.

[2] Chirath Kumarasiri’s and Cassim Faroo’s focuses on a Part-of-Speech (POS) Tagger based NLP technique for aspect identification from reviews. Then a Naïve Bayes (NB) Classifier is used to classify identified aspects into meaningful categories.

[3] I. K. C. U. Perera and H.A. Caldera have used data mining techniques like Opinion mining and Sentiment analysis to automate the analysis and extraction of opinions in restaurant reviews.

[4] Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga and Akila Pemasiri wrote a paper on ‘Eatery, a multi-aspect restaurant rating system’ that identifies rating values for different aspects of a restaurant by means of aspect-level sentiment analysis. This research introduced a new taxonomy to the restaurant domain that captures the hierarchical relationships among entities and aspects.

[5] Neha Joshi wrote a paper in 2012 on A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City which aims to contribute to the limited research in this area and provide
insight into the consumer decision making process specifically for the India foodservice industry. She did hypothesis testing using chi-square test.

[6] Bidisha Das Baksi, Harrsha P, Medha, Mohinishree Asthana, Dr. Anitha C wrote a paper that studies various attributes of existing restaurants and analyses them to predict an appropriate location for higher success rate of the new restaurant. The study of existing restaurants in a particular location and the growth rate of that location is important prior to selection of the optimal location. The aim is to the create a web application that determines the location suitable to establish a new restaurant unit, using machine learning and data mining techniques.

3. DATA SET DESCRIPTION

This is a kaggle dataset.
(https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants).

It Represents information of Restaurants in the City of Bangalore.

It contains 17Columns and 51,000 Rows

3.1 PreProcessing

The Dataset contained 17 Attributes.

- Records with null values were dropped from ratings columns and were replaced in the other columns with a numerical value.
- Values in the 'Rating' column were changed. The '/5' string was deleted. For eg. If the rating of a restaurant was 3.5/5, it was changed to 3.5.
- Using LabelEncoding from sklearn library, encoding was done on columns like book_table,online_order,rest_type,listed_in(city).

3.2 Feature Selection

We did not use any feature selection algorithms but eliminated some columns due to available domain knowledge and thorough study of the system.

Dropped columns mentioned below:

- URL
- Address
- Dish_liked
- Phone
- Menu
- Review_list
- Location
- Cuisine

Some of these columns may look like they are important but all of the same information could be found in other columns with lesser complexity.

The Columns being used are as follows:

- Name
- Online_order
- Book_table
- Votes
- Rest_type
- Approx. cost of two people
- Listed_in(type)
- Listed_in(city)

4. EXPLORATORY DATA ANALYSIS

A lot of effort went into the EDA as it gives us a detailed knowledge of our data.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

1) Restaurant Rate Distribution

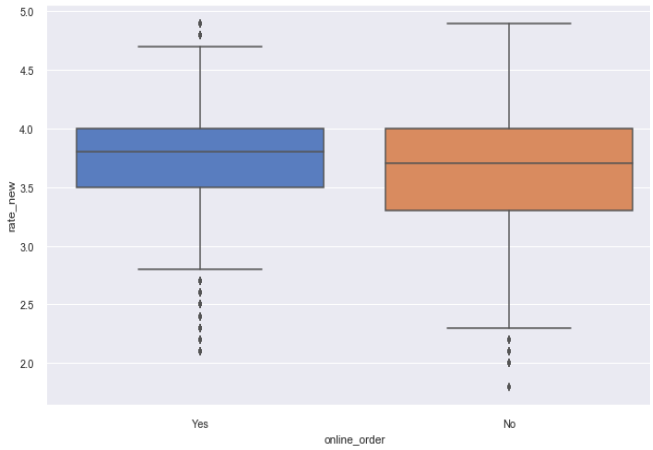
We can see that the number of restaurants with the rating between 3.5 and 4 are the highest. We will look into its dependencies further.

2) Approximate Cost of two people

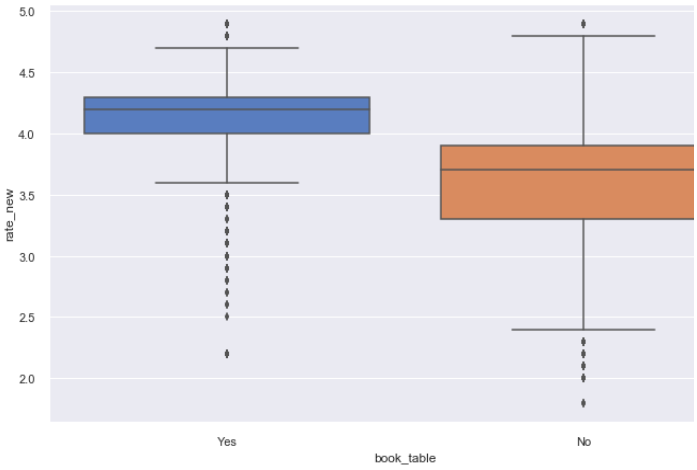
This is a graph for the 'Approximate cost of 2 people' for dining in a restaurant. Restaurants with this cost below 1000 Rupees are more.

This box plot helps us look into the outliers. We can also see that online ordering service also affects the rating. Restaurants with online ordering service have a rating from 3.5 to 4.

3) Online ordering with respect to Rating(Finding Outliers)

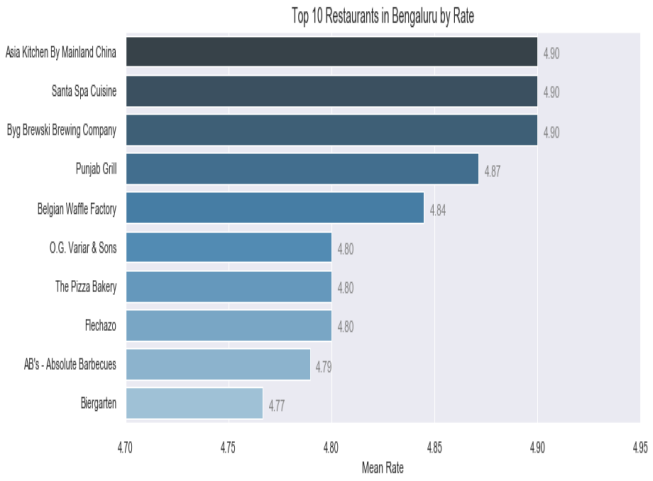


4) Booking table with respect to rating(Finding Outliers)



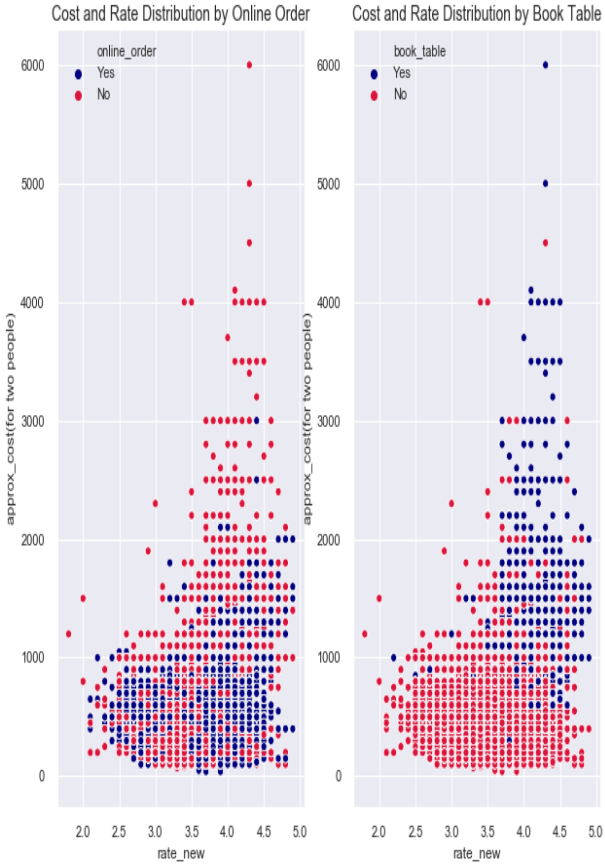
This box plot also helps us look into the outliers. This box plot is regarding how table booking availability is seen in restaurants with rating over 4.

5) Top Rated Restaurants



This graph just showcases the best restaurants in Bangalore along with their rating.

6) Cost and Rate Distribution according to online ordering and booking table



A very important scatterplot shows the correspondence between the cost, online ordering, bookings and rating of the restaurant.

4.1. Key Findings

	Votes	approx_cost(for two people)	Rating
online_order			
No	367.992471	716.025190	3.658071
Yes	343.228663	544.365434	3.722440

	Votes	approx_cost(for two people)	Rating
Book_table			
No	204.580566	482.404625	3.620801
Yes	1171.342957	1276.491117	4.143464

5. RESULTS

Algorithms	Accuracy
Linear Regression	30%
KNN	44%
Support Vector Machine	43%
Decision Tree	69%
Random Forest	81%
ADA Boost(DT)	83%
XGBoost	72.26%
Gradient Boosting	52%

In this model, we have considered various restaurants records with features like the name, average cost, locality, whether it accepts online order, can we book a table, type of restaurant.
This model will help business owners predict their rating on the parameters considered in our model and improve the customer experience.
Different algorithms were used but in the end the final model is selected on Ada Boost Regressor which gives the highest accuracy compared to others.

6. CONCLUSIONS

This paper studies a number of features about existing restaurants of different areas in a city and analyses them to predict rating of the restaurant. This makes it an important aspect to be considered, before making a dining decision. Such analysis is essential part of planning before establishing a venture like that of a restaurant.
Lot of researches have been made on factors which affect sales and market in restaurant industry. Various dine-scape factors have been analysed to improve customer satisfaction levels.
If the data for other citirs is also collected, such predictions could be made for accurate.

7. REFERENCES

[1] Chirath Kumarasiri, Cassim Faroo,"User Centric Mobile Based Decision-Making System Using Natural Language Processing (NLP) and Aspect Based Opinion Mining (ABOM) Techniques for Restaurant Selection". Springer 2018. DOI: 10.1007/978-3-030-01174-1_4

[2] Shina, Sharma, S. & Singha ,A. (2018). A study of tree based machine learning Machine Learning Techniques for Restaurant review. 2018 4th International Conference on Computing Communication and Automation (ICCCA)
DOI:/10.1109/CCAA.2018.8777649

[3] I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546. doi: 10.1109/CIAPP.2017.8167276

[4] Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga and Akila Pemasiri. Eatery – A Multi-Aspect Restaurant Rating System. Conference: the 28th ACM Conference

[5] Neha Joshi. A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City.
Global Journal of Management and Business Research(2012)
Link:
https://pdfs.semanticscholar.org/fef5/88622c39ef76dd773fcad8bb5d233420a270.pdf

[6] Bidisha Das Baksi, Harrsha P, Medha, Mohinishree Asthana, Dr. Anitha C.(2018) Restaurant Market Analysis.
International Research Journal of Engineering and Technology (IRJET)
Link: https://www.irjet.net/archives/V5/i5/IRJET-V5I5489.pdf

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 379-384, 2019, ISSN:-2319–8656

Engineering College Admission Preferences Based on Student Performance

Dhruvesh Kalathiya
Student, M.Sc.(BDA)
MIT-WPU
Pune, India

Rashmi Padalkar
Student, M.Sc.(BDA)
MIT-WPU
Pune, India

Rushabh Shah
Student, M.Sc.(BDA)
MIT-WPU
Pune, India

Sachin Bhoite
Assistant Professor
Department of Computer Science
Faculty of Science
MIT-WPU
Pune, India

Abstract: As we know that after the 12th board results, the main problem of a student is to find an appropriate college for their further education. It is a tough decision to make for many students as to which college they should apply to. We have built a system that compares the student's data with the past admission data and suggests colleges in a sequence of their preference. We have used Decision Tree, Support Vector Classifier, Extra Tree Classifier, Naïve Bayes, KNN and Random Forest as our statistical model to predict the probability of getting admission to a college. It was observed that the performance of Random Forest was achieved highest among all.

Keywords: Decision Tree, Random Forest, KNN, Random Forest, Extra Tree Classifier, SVC, Probabilities

1. INTRODUCTION

Education plays a vital role in today's era. While we talk about career – a person's degree, course, university and the knowledge that he possesses – is the key factor on which the firm hires a fresher. As soon as a student completes his/her Higher Secondary Schooling, the first goal of any student is to get into an appropriate College so that he can get a better education and guidance for his future. For that, students seek help from many sources like online sites or career experts to get the best options for their future. A good career counselor charges a huge amount for providing such solutions. Online sources are also not as reliable as the data from a particular source is not always accurate. Students also perform their analysis before applying to any institution, but this method is slow and certainly not consistent for getting actual results and possibly includes human error. Since the number of applications in different universities for each year is way too high, there is a need to build up a system that is more accurate or precise to provide proper suggestions to students.

Our aim is to use machine learning concepts to predict the probability of a student to get admission into those preferred colleges and suggest a list of colleges in a sequence of the probability of getting admission to that specific college. The following are the steps that include the work we have done in sequence of implementation.

2. RELATED WORKS

One of the researchers has done work on predicting the university and students applying to explicit universities (Jay Bibodi).The first one is the University selection model, and the second one is a student selection model. They came across some issues like noisy data, unformatted text but after cleaning the data, they proceeded to 'model selection' with some important features. "University Selection Model" – A Classification problem with apriori probability output. They found out just two universities giving a higher probability of output. "Student Selection Model" – Classification using supervised learning like Linear and kernel, Decision Tree and Random Forest. Random Forest provided better accuracy than other algorithms i.e. 90% accuracy.[1]

There is one more researcher – Himanshu Sonawane – who has researched on 'Student Admission Predictor'. It is a system built to help students who are studying abroad. This system helps students find the best foreign universities/colleges based on their performance in GRE, IELTS, 12^{th}, Graduation Marks, Student Statement of purpose, Letter of Recommendation, etc. Based on this information, it recommends the best-suited university/college. They have used three algorithms: KNN (76% Accuracy), Decision Tree (80% Accuracy), Logistic Regression (68% Accuracy). In the case of a decision tree, accuracy was nearly the same for both pieces of training as well as testing datasets.[2]

From another research paper, we got to know what affects the likelihood of enrolling (Ahmad Slim – Predicting Student Enrolment Based on Student and College Characteristics). They have used machine learning to analyze the enrolment.

This work intends to provide decision-makers in the enrolment management administration, a better understanding of the factors that are highly correlated to the enrolment process. They have used real data of the applicants who were admitted to the University of New Mexico (UNM).In their dataset, they have different features like gender, GPA, parent's income, student's income. They had data issues like missing value and categorical variables. They have divided classification into two parts – classification at the individual level and classification at a cohort level. For classification at the individual level, the model was used to check the probability of enrolment and whether the applicant is enrolled or not. Logistic Regression (LR) provided an accuracy of 89% and Support Vector Machine (SVM) provided an accuracy of 91% which was used in the classification at an individual level. The total enrolment in 2016 was actually 3402 but the prediction was 3478 by using past year records (2015) using time series for classification at the cohort level. [3]

These researchers – Heena, Mayur, and Prashant from Mumbai – have used data mining and ML techniques to analyze the current scenario of admission by predicting the enrolment behavior of students. They have used the Apriori technique to analyze the behavior of students who are seeking admission to a particular college. They have also used the Naïve Bayes algorithm which will help students to choose the course and help them in the admission procedure. In their project, they were conducting a test for students who were seeking admissions and then based on their performance, they were suggesting students a course branch using Naïve Bayes Algorithm.[4]

One more researcher has made a project for helping students in suggesting them best-suited colleges in the USA based on his/her profile. He has collected the data from online sources which was reported by students. He has used 5-6 algorithms for his project. Naïve Bayes was one of them which gave the highest accuracy among all of them. He has predicted students' chances(probabilities) of getting admission in 5 different universities in the USA.[5]

Other researchers were predicting the student admission (Students' Admission Prediction using GRBST with Distributed Data Mining - Dinesh Kumar B Vaghela). They have used the Global Rule Binary Search Tree (GRBST). While searching, they identified some problems like maintaining a single database for all the colleges were difficult. This paper has two phases i.e. training phase and testing phase. In the training phase, the J48 algorithm was used for all local sites. In the testing phase, Users can interact with the system with the help of the application layer. They have used consolidation techniques in two ways i.e. using If…Then… rules format and Decision Table. They have also used binary search tree construction. After applying this technique, they have found the time complexity of generating the Binary Search Tree from the Decision table is very less and also this BST has efficient time complexity to predict the result. They conclude that data mining techniques can be useful in deriving patterns to improve the education system. [6]

GRADE system developed to help graduate admission committee at the University of Texas at Austin Department of Computer Science (UTCS) by Austin Waters and Risto Miikkulainen Department of Computer Science 1 University Station C0500, University of Texas, Austin, TX 78712. This system first reads the applicant's files from the database and encodes as a high-dimensional feature vector and then a logistic regression classifier is trained on that data. It then predicts the probability of binary classification. The feature vector encoding of a student's file indicates whether the applicant was rejected or admitted. The system was used to predict the probability of admissions committee accepting that applicant or not but, in our model, we are trying to make it easy for the applicants to understand whether they should apply to that college or not.[7]

3. DATA EXTRACTION AND TRANSFORMATION

We have achieved our goals step-by-step to make the data steady, fitting it into our models and finding out suitable algorithms of machine learning for our System.

This step contains mainly – Data Extraction, Data Cleaning, Pre-processing, removing unwanted columns, feature selection, label encoding. These steps are shown in Figure 1.

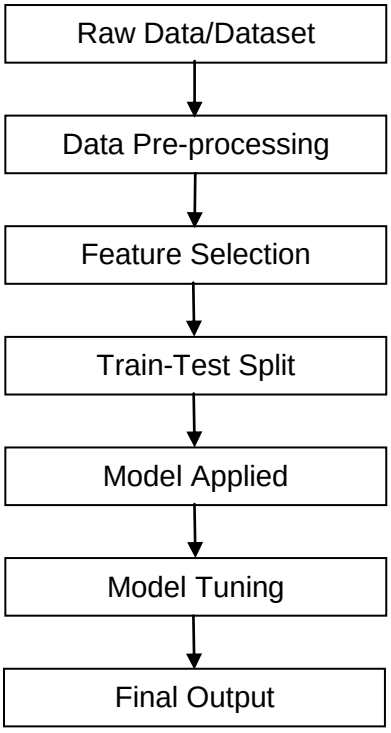


Figure 1: Architecture

3.1 Dataset

Knowing about this use case we need past admission data of multiple colleges to work on. We have extracted data from three different colleges which includes information about a student's academic scores and the reservation category he falls in. Data has been mined from college registries. We have extracted 2054 records that include 13 attributes.

Attributes of the dataset are:

First Name, Last Name, Email ID, Gender, Address, Date of Birth, Category, S.S.C. Percentage, H.S.C Percentage, Diploma Percentage, Branch, Education Gap, and Nationality.

1. Data Preprocessing

Data preprocessing is an important task in machine learning. It converts raw data into clean data. Following are techniques, we have applied on data: -

- **Missing Values –** Missing Value are those values that failed to load information or the data itself was corrupted. There are different techniques to handle missing values. One of which we have applied is deleting rows because some of the rows were blank and they may mislead the classification.
- **Label Encoder –** This is one of the most frequently used techniques for the categorical variable. Label encoder converts labels into a numeric format so that the machine can recognize it. In our data, there are many attributes which are categorical variable like gender, category, branch.
- **Change in data type –** Some attributes didn't include proper input. For example, the Nationality attribute included values like Indian, India, IND which all meant the same country. For that purpose, we needed to change such values into a single format. 'Object' data type values in some attributes had to be changed into 'float' data type. Some records included CGPA for S.S.C scores so we converted those records into a percentage. We made all these changes so that it doesn't affect our accuracy.
- **Drop Columns –** As per domain knowledge, we removed some columns which were not needed in our model.

2. Feature Selection

As we proceed further, before fitting our model we must make sure that all the features that we have selected contribute to the model properly and weights assigned to it are good enough so that our model gives satisfactory accuracy. For that, we have used 4 feature selection techniques: Lasso, Ridge, F1 Score, Extra Tree Classifier.

Lasso, Ridge and F1 Score were removing the features that I needed the most and Extra Tree Classifier was giving me an acceptable importance score. Which is shown below.

Extra Tree Classifier:

Extra Tree Classifier is used to fit a randomized decision tree and uses averaging to improve the predictive accuracy and control over-fitting. We have used this to know the important features of our data.

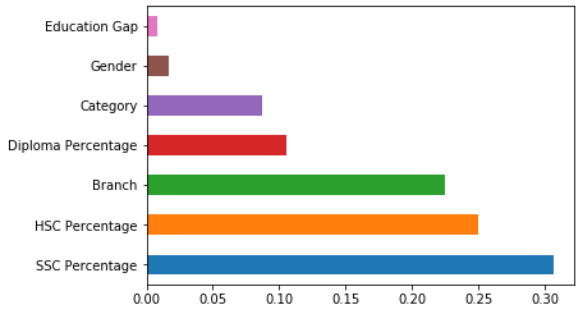


Figure 2: Feature Selection using Extra Tree Classifier

As we can see, my Feature Selection model is giving more importance to S.S.C. The percentage is not appropriate.

So, in this case, our domain knowledge is also helpful to make decisions for this type of situation.

4. EXPLORATORY DATA ANALYSIS

As we saw in feature selection, some features which seemed not so important were contributing to our model. So, to understand those features, we need to do exploratory analysis on this data.

We did exploratory analysis on a few features by grouping and plotting it on graphs.

EDA on Gender Column:

By grouping the gender and plotting the admissions in different colleges as per their gender, we identified some relations between the student's admission and his or her gender. As shown in Figure 4 – For different gender, most students lied in different bins for different colleges. Even for different colleges, we are getting different bell curves. Looking at this we can confirm that the gender column is contributing to our model.

For Extra Tree Classifier, Gender contributes to model – 1.3092%.

EDA on Category Column:

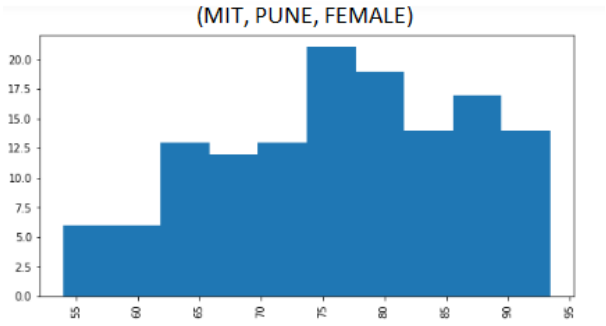
By grouping the category and calculating the percentage of students who got admissions with respect to their categories is shown in Figure 3 – For different categories, we calculated the percentage of students that lie in each category. This percentage of students was matching to reservation criteria as per Indian laws. This shows that the Category column is contributing to our model.

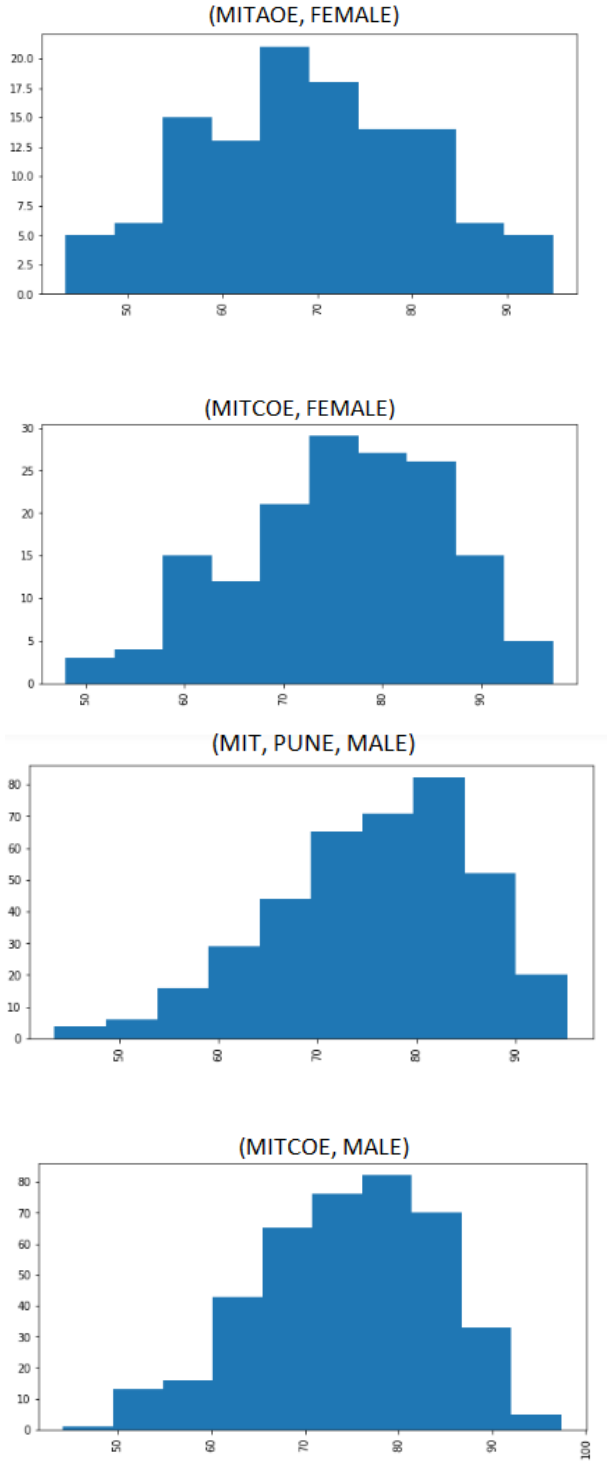
For Extra Tree Classifier, Category contributes to model – 9.6582%.

```
In [5]: df.groupby("Category")["Sr. no."].count()/df.shape[0]*100

Out[5]: Category
        NT(B)      1.850950
        NT(C)      3.214808
        NT(D)      1.899659
        OBC       19.629810
        OPEN      55.577204
        SBC        1.802241
        SC         8.426693
        ST         2.143205
        VJ(A)      1.802241
        select     3.604481
        Name: Sr. no., dtype: float64
```

Figure 3: Admission with respect to Category





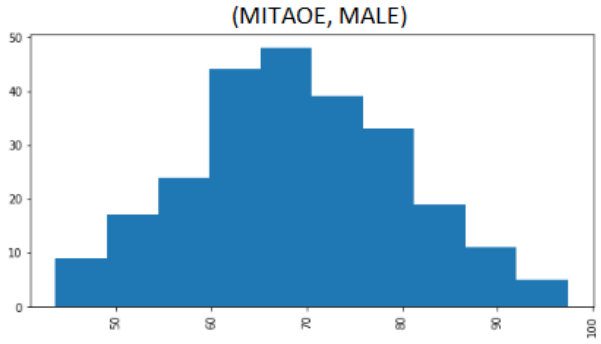


Figure 4: Admission with respect to Gender

5. RESULT AND DISCUSSION

After removing all the noise from the data and after selecting appropriate features for our model, the next step is to find out the best model which gives us more accuracy for train and test both. But before that, we must split our data into 2 parts as we don't have any testing dataset right now.
So, we have divided this modeling section into 3 parts:

1. Train-Test Split
2. Modeling
3. Tuning Model

5.1 Train-Test Split

The training data set is used to create the model while testing the data set is used to qualify the performance. Training data's output is available to model while test data is unseen data. So, in our data, we have split data into 70% for training data and 30% for testing data because it makes the classification model better. while the test data makes the error estimate more accurate.

5.2 Modeling

Following are models, we have applied to check which model gives better accuracy:

- **Support Vector Classifier (SVC):**
  This algorithm is used for the classification problem. The main objective of SVC is to fit the data you provide, returning a "best fit" hyperplane that divides or categorizes your data. From there, when obtaining the hyperplane, you'll then feed some options to your category to examine what the "predicted" class is.

- **Decision Tree:**
  A decision tree is non-parametric supervised learning. It is used for both classification and regression problems. It is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path between root and leaf represents classification rules. It creates a comprehensive analysis along with each branch and identifies decision nodes that need further analysis.

- **Random Forest:**
  Random Forest is a meta estimator that uses the number of decision trees to fit the various subsamples

drawn from the original dataset. we can also draw the data with replacement as per the requirements.

- **K-Nearest Neighbors (KNN):**
  K-Nearest Neighbors (KNN) is a supervised learning algorithm that is used to solve regression as well as classification problems. Where 'K' is the number of nearest neighbors around the query. It is simple to implement, easy to understand and it is a lazy algorithm. The lazy algorithm means it does not need any training data points for a model generation [8]. All the training data is used in the testing phase. This makes the training faster and the testing phase slower and costlier. By costly testing phase we mean it requires more time and more memory.

- **Naïve Bayes:**
  A Naive Bayes Classifier is a supervised machine-learning algorithm that uses Bayes' Theorem, in which the features are statistically independent. It specifies multiple uses of probability theories and statistics. By simple machine learning problem, where we need to teach our model from a given set of attributes (in training examples) and then form a hypothesis or a relation to a response variable. Then we tend to use this to predict a response, given attributes of a replacement instance.

- **Extra Tree Classifier:**
  The main objective of Extra Tree Classifier is randomizing tree building further in the context of numerical input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree.

We used all these models to fit our data and checked the accuracy which is shown in Figure 5.

Model	Accuracy
Support Vector Classifier (SVC)	47.79%
Decision Tree	56.13%
Random Forest Tree	58.87%
K-Nearest Neighbors (KNN)	52.35%
Naïve Bayes	42.29%
Extra Tree Classifier	58.33%

Figure 5: Accuracy for Different Models

5.3 Model Tuning:

As we can see our accuracy is not going beyond 60%. for that reason, we have tuned our model. Tuning is the method for increasing a model's performance without overfitting the data or making the variance too high. Hyperparameters disagree from other model parameters therein they're not learned by the model automatically through training ways. Following are the models, we applied to check which model gives better accuracy:

- **XGBoost:**
  XGBoost stands for eXtreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance [9]. Using this we have achieved 64% accuracy.
- **AdaBoost:**
  AdaBoost is one of the first boosting algorithms to be adapted in solving practices. AdaBoost helps you combine multiple "weak classifiers" into one "strong classifier". AdaBoost is best used to boost the performance of all trees on binary classification issues. Using this we have achieved 61% accuracy.

We have used XGBoost and AdaBoost for just improving our accuracy. Accuracy of XGBoost is higher and it improves our accuracy by 6%.

But as our problem statement suggests, we do not need accuracy as we are just calculating the probabilities for getting the admission in all the colleges and referring top probabilities to that student.

6. CONCLUSIONS

The objective of this project is achieved in this process flow which will be used by students to identify the appropriate colleges based on his/her performance. The main aspects of students which are taken under are their 10^{th}, 12^{th} percentages and diploma percentage too if applicable. Besides of that gender, category, Education gap and branch in which student wants to get admission are also contributing to admission. The final model for our project is Random Forest as it is giving a satisfactory output.

As we looked at our data, and we observed that this is just the data of students who took admission. There is no data of neither rejected students nor we have students' choice of college. We can ask students for their choice for college to get better data for accuracy. We were also about to consider the address of the student, as we know that different seats are reserved for those students who belong to different states. But in our data, most of the address columns are not filled properly so we removed that column. So, for that, we can keep dropdown buttons on Online Application Form for cities, states, and countries so that we get proper data for this. We also can ask for entrance exam results which can help us predict more accurately.

7. REFERENCES

[1] Bibodi, J., Vadodaria, A., Rawat, A. and Patel, J. (n.d.). "Admission Prediction System Using Machine Learning". California State University, Sacramento.

[2] Himanshu Sonawane, Mr. Pierpaolo Dondio. "Student Admission Predictor". School of Computing, National College of Ireland. unpublished.

[3] A. Slim, D. Hush, T. Ojah, T. Babbitt. [EDM-2018] "Predicting Student Enrollment Based on Student and College Characteristics". University of New Mexico, Albuquerque, USA.

[4] Heena Sabnani, Mayur More, Prashant Kudale." Prediction of Student Enrolment Using Data Mining Techniques". Dept. of Computer Engineering, Terna Engineering College, Maharashtra, India.

[5] Bhavya Ghai. “Analysis & Prediction of American Graduate Admissions Process”. Department of Computer Science, Stony Brook University, Stony Brook, New York.

[6] Dineshkumar B Vaghela, Priyanka Sharma. “Students' Admission Prediction using GRBST with Distributed Data Mining”. Gujarat Technological University, Chandkheda.

[7] Austin Waters, Risto Miikkulainen. “GRADE: Machine Learning Support for Graduate Admissions”. University of Texas, Austin, Texas.

[8] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[9] https://www.meetup.com/Big-Data-Analytics-and-Machine-Learning/events/257926117/

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 385-388, 2019, ISSN:-2319–8656

Wine Quality Prediction using Machine Learning Algorithms

Devika Pawar[1]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Aakanksha Mahajan[2]
M.Sc. (Big Data Analytics)
MIT-WPU
Pune, India

Sachin Bhoithe[3]
Faculty of Science
MIT-WPU
Pune, India

Abstract: Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality. Classification models used here are 1) Random Forest 2) Stochastic Gradient Descent 3) SVC 4)Logistic Regression.

Keywords: Machine Learning, Classification,Random Forest, SVM,Prediction.

I. INTRODUCTION

The aim of this project is to predict the quality of wine on a scale of 0–10 given a set of features as inputs. The dataset used is Wine Quality Data set from UCI Machine Learning Repository. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol. And the output variable is quality (score between 0 and 10).We are dealing only with red wine. We have quality being one of these values: [3, 4, 5, 6, 7, 8]. The higher the value the better the quality. In this project we will treat each class of the wine separately and their aim is to be able and find decision boundaries that work well for new unseen data. These are the classifiers.

In this paper we are explaining the steps we followed to build our models for predicting the quality of red wine in a simple non-technical way. We are dealing only with red wine. We would follow similar process for white wine or we could even mix them together and include a binary attribute red/white, but our domain knowledge about wines suggests that we shouldn't. Classification is used to classify the wine as good or bad. Before examining the data it is often referred to as supervised learning because the classes are determined.

II. RELATED WORK

Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing.

College of Intelligent Science and Engineering, China has written a paper on Evaluation and Analysis Model of Wine Quality Based on Mathematical Model.They have used various mathematical test to predict the quality of wine.The Mann-Whitney U test is used to analyze the wine evaluation results of the two wine tasters, and it is found that the significant difference between the two is small. Then this paper uses the Cronbach Alpha coefficient method to analyze the credibility of the two groups of data.[1]

Paulo Cortez ,Juliana Teixeira,António CerdeiraFernando AlmeidaTelmo MatosJosé Reis wrote a paper on wine Quality assesment using Data Mining techniques.In this paper,they proposed a data mining approach to predict wine preferences that is based on easily available analytical tests at the certification step. A large dataset was considered with white vinho verde samples from the Minho region of Portugal. Wine quality is modeled under a regression approach, which preserves the order of the grades. 95% accuracy was obtained using these data mining techniques.[2]

The study of this paper was done at International Journal of Intelligent Systems and Applications in Engineering and this paper was published on 3rd September 2016. The main objective of this research paper was to predict wine quality based on physicochemical data. In this study, two large separate data sets which were taken from UC Irvine Machine Learning Repository were used. The instances were successfully classified as red wine and white wine with the accuracy of 99.5229% by using Random Forests Algorithm.[3]

III. PROPOSED WORK

A. Data Set:

Dataset/Source: **Kaggle**
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Structured/Unstructured data: Structured Data in CSV format.

Dataset Description: The two datasets are related to red wine of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

1)fixed acidity
2) volatile acidity
3) citric acid
4) residual sugar
5) chlorides
6)free sulfur dioxide
7)total sulfur dioxide
8)density
9)pH
10) sulphates
11) alcohol
Output variable (based on sensory data):
12)quality (score between 0 and 10)

IV. DATA PROCESSING METHODS

For making automated decisions on model selection we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, we are using F1 score which combines two metrics: Precision which expresses how accurate the model was on predicting a certain class and Recall which expresses the inverse of the regret of missing out instances which are misclassified. Since we have multiple classes we have multiple F1 scores. We will be using the unweighted mean of the F1 scores for our final scoring. This is a business decision because we want our models to get optimized to classify instances that belong to the minority side, such as wine quality of 3 or 8 equally well with the rest of the qualities that are represented in a larger number. For the regression task we are scoring based on the coefficient of determination, which is basically a measurement of whether the predictions and the actual values are highly correlated. The larger this coefficient the better. For regressors we can also get F1 score if we first round our prediction.

Splitting for Testing : We are keeping 20% of our dataset to treat it as unseen data and be able and test the performance of our models. We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset. Other than that the selection is being done randomly with uniform distribution.

Various classification and regression algorithms are used to fit the model. The algorithms used in this paper are as follows:

For classification:

Random Forest Decision Trees classifier

Support Vector Machine classifier

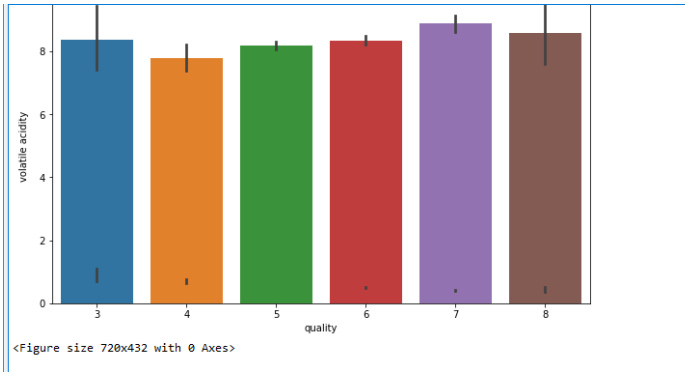
Stochastic gradient descent

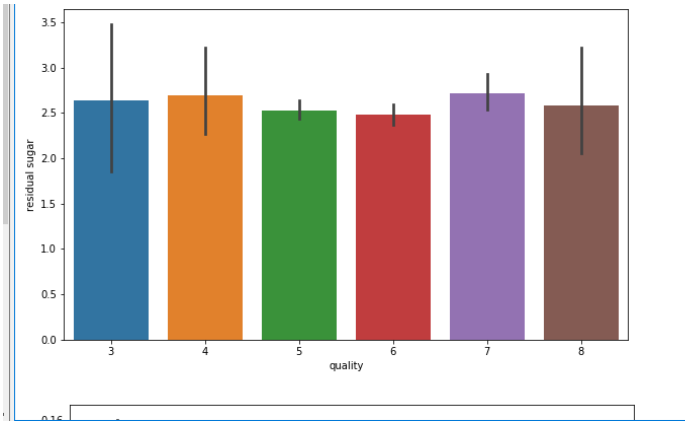
Logistic Regression classifier

Preprocessing: Label Encoding is used to convert the labels into numeric form so as to convert it into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning. We have used label encoding to label the quality of data as good or bad. Assigning 1 to good and 0 to bad.

Feature Selection:

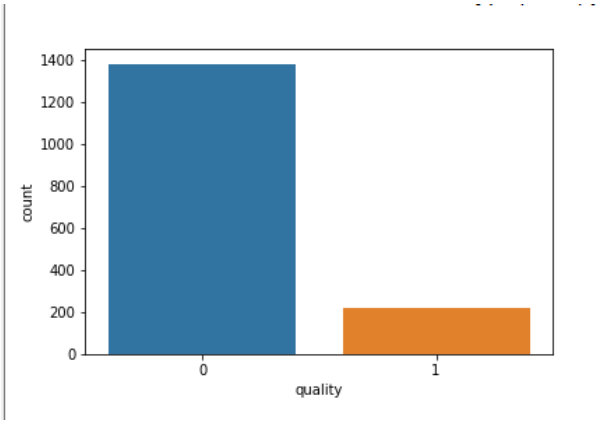
As we can clearly see, volatile acidity and residual sugar are both not very impact full of the quality of wine. Hence we can eliminate these features. Though we are selecting these features, they will change according to the domain experts.



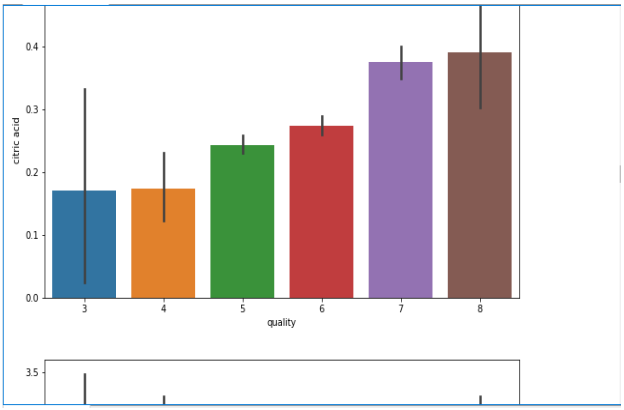


Exploratory Data Analysis:

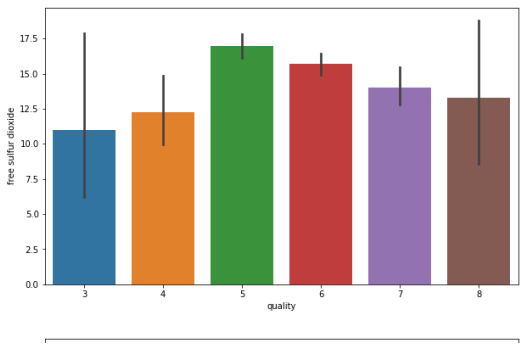
- The below bar plot shows the count of data which is good or bad. We can see 80% of the data is classified with good wine quality and 20% with bad quality of wine.



- This bar plot shows a directly proportional relation between citric acid and quality.As the quality of wine increases the amount of citric acid also increases which shows that citric acid is the important feature on which quality of wine depends.



- Free sulphur dioxide is greatly contributing to the quality of wine, this bar plot gives us a more clear picture.



Result and Discussion: Algorithms used for classification are:

1) Logistic Regression
2) Stochastic gradient descent
3) Support Vector Classifier
4) Random Forest

- Logistic Regression gave us an accuracy of 86%

  Performance matrix of Logistic Regression:

	Precision	Recall	F1-Score	Support
0	0.88	0.98	0.93	273
1	0.71	0.26	0.37	47

- Stochastic gradient descent was able to give an average accuracy of 81%.
  Performance matrix of SGD:

	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	273
1	0.42	0.30	0.35	47

- Support Vector Classifier has given an accuracy of 85%.

  Performance matrix of SVC:

	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	273
1	0.71	0.26	0.37	47

- Random Forest gave us an accuracy of 87.33%

	Precision	Recall	F1-Score	Support
0	0.90	0.97	0.93	273
1	0.68	0.40	0.51	47

CONCLUSION

Based on the bar plots plotted we come to an conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not have change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection.

For classifying the wine quality, we have implemented multiple algorithms, namely

1) Logistic Regression

2) Stochastic gradient descent

3) Support Vector Classifier

4) Random Forest

We were able to achieve maximum accuracy using random forest of 88%. Stochastic gradient descent giving an accuracy of 81% .SVC has an accuracy of 85% and logistic regression of 86%.

References:

[1] Yunhui Zeng1 , Yingxia Liu1 , Lubin Wu1 , Hanjiang Dong1. “Evaluation and Analysis Model of Wine Quality Based on Mathematical Model ISSN 2330-2038 E-ISSN 2330-2046,Jinan University, Zhuhai,China.

[2] Paulo Cortez1, Juliana Teixeira1, Ant´onio Cerdeira2.“Using Data Mining for Wine Quality Assessment”.

[3] Yesim Er*1 , Ayten Atasoy1. “The Classification of White Wine and Red Wine According to Their Physicochemical Qualities”,ISSN 2147-67992147-6799,3rd September 2016

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 389-393, 2019, ISSN:-2319–8656

Expert System for Student Placement Prediction

Krishna Gandhi
Msc. Data Science and Big Data Analytics
MIT- WPU
Kothrud,Pune, Maharashtra, India.

Aadesh Dalvi
Msc. Data Science and Big Data Analytics
MIT- WPU
Kothrud, Pune, Maharashtra, India.

Aniket Walse
Msc. Data Science and Big Data Analytics
MIT- WPU
Kothrud, Pune, Maharashtra, India

Sachin Bhoite
Msc. Data Science and Big Data Analytics
MIT- WPU
Kothrud, Pune, Maharashtra, India

Abstract: - Data mining is a process of extracting and identifying previously unknown and potentially useful information or pattern from large amount of data using different methods and techniques. Data mining in a domain of education is known as Educational data mining (EDM). This paper discusses about an expertise system which can used as student placement prediction system. A statistical model is applied on a reputed college's past data after data pre-processing and feature selection. This model can be used to predict the percentage of chances of a student getting selected in campus placement. It will help students evaluating themselves and identifying which skills are essential.

Keywords:- Data Mining, Educational Data Mining, Expertise system, Statistical model, Data pre-processing, Feature selection.

1. INTRODUCTION

This model is about concerning those students who wants to get a better placement for better future. Sometimes it happens that student gets sidetracked from studies in initial Semesters and later they realize the importance of marks/CGPA. Basically, this will help them to enhance their performance and will make them believe that they can achieve their dream job. Application of EDM is an evolving trend in the worldwide [1]. This will help college faculty to show the precise roadmap to students when it comes to placement and choose their career path.

It will guide colleges and institutions to maintain their reputation by making most of the placement. It drives students to ask questions regarding what can nurture them. It can give an overview to

Junior college students while selecting their stream.

1) What subjects to target?
2) What skills to improvise on?
3) Probability of getting placed after choosing their specialization?

Data visualization will help College students to get a clearer view regarding which stream they should choose. This can be done using different libraries of Python like Matplotlib, where student and faculties can visualize overview of each stream.

This paper describes the model to predict the percentage of skills required by engineering students pursuing Bachelors and Master's Degree with respect to company's skillset necessity. According to the rules generated, percentage of selection of students will vary. These rules are generated with the help of Domain Expert.

Percentage of Selection = (Criteria's Satisfied/Number of Criteria's) * 100

1.1 Literature Review

The researchers have studied several related national & international research papers, thesis to understand aims, technique used, various expert systems, datasets, data preprocessing approaches, features selection methods, etc.

Siddu P. Algur, Prashant Bhat and Nitin Kulkarni used two algorithms- Random Tree and J48 to construct a classification models using Decision Tree concept. The Random Tree classification model is more effective as compared to J48 classification model [2].

Machine learning algorithms are applied in weka environment and R studio by K. Sreenivasa Rao, N. Swapna and P. Praveen Kumar. Results is tabulated and analyzed, It shows random tree algorithm gives 100% accuracy in prediction on their dataset and also in R environment Recursive Partitioning & Regression Tree performs better and gives 90% accuracy. We also accept that performance depends on nature of dataset [3].

V.Ramesh, P. Parkavi and P. Yasodha also proved that Multilayer Perception algorithm is most suitable for predicting student performance. MLP gives 87% prediction which is comparatively higher than other algorithms [4].

K. Sripath Roy 1, K. Roopkanth, V. Uday Teja, V. Bhavana, J. Priyanka. The data is trained and tested with all three algorithms and out of all SVM gave more accuracy with 90.3% and then the XG Boost with 88.33% accuracy [5].

Ajay Kumar Pal, met with his goal and proved that the top algorithm is Naïve Bayes Classification with an accuracy of 86.15% with an error average of 0.28 with others. He also conveyed that naïve Bayes has the potential to classify conventional methods [6].

Sudheep Elayidom, Dr. Suman Mary Idikkula and Joseph Alexander, studying past data and and following the trend, and based on that the judgment for future will be given.

2. DATASET DESCRIPTION

The data used in this model is supplied by a well-known Engineering College situated in Pune, Maharashtra. Data generated is collected from the details given by graduates, post graduates, diploma holders in engineering of various streams during the year 2019. It includes students 10^{th}, 12^{th} or Diploma and semester-wise aggregation for Bachelors and Master's. Dataset contains 2330 tuples and 81 attributes holding multiple streamwise data of the students.

2.1 Data Pre-Processing

Data has redundant, incomplete, inconsistent and inaccurate entries. We discovered that there were many different attributes which seems to be superfluous and which won't affect our results. By consulting our Domain Expert, we decided to remove those attributes as well as tuples using tools like excel.
Entries with human errors seem to be illusory. So as per discussion with Expert we decide to apply mean to the data using Python.

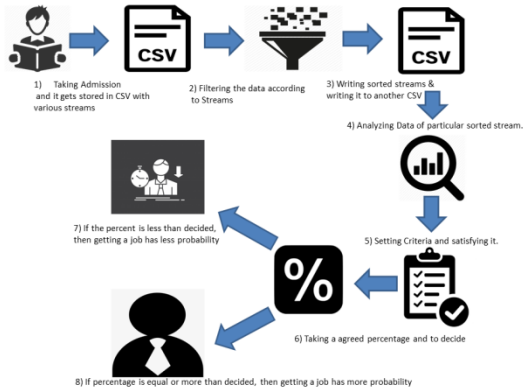
2.2 Feature Selection

Attributes impacting the placements of students were taken into consideration with the help of Expert advice.

Factors like 10th, 12th or Diploma and Degree Aggregation are too affecting the placement predictions for students as well as their non-educational attributes like work-experience, projects and external certification were also taken into consideration.

3. PROPOSED EXPERT SYST-EM

After analyzing a lot we came to an idea of creating our own system, we will help us to understand the genuine accuracies for placement.



1) A student fills the form and gets stored in CSV.
2) CSV contains all details of all the streams data of all students
3) Then system filters all streams and writes one by one detail in each new CSV.
4) Then domain expert analyzes it on each stream and sets criteria for all the attributes.
5)The system checks whether students get satisfied or not according to the specific criteria set.
6) Then the criteria will decide which students are going to be placed in a campus placement and which are not.

3.1 Pseudo Code Of The System

Pseudo is basically the demo code which we have decided to implement on our model. It is type of rule setting methodology where student have to fulfill all criteria's for getting a reputed placements.

1. Load all student data.
2. Form rule based on stream with the help of expert
3. Enter the data of new students for the prediction of the placement.
4. Calculate Number of criteria satisfied by that student.

```
        Flag=0
If rule 1 satisfied
        Flag=flag+1
        Else
        Flag=flag
        If rule 2 satisfied
        Flag=Flag+1
        Else
            Flag=flag
```

5. Perform this operation for all the criteria created.
6. Calculate our prediction by

   Prediction percentage= (Flag/Number of rules) *100

4. EXPLORATORY ANAYLISIS

A pie chart is a circular statistical illustration, which is divided into different parts to demonstrate numerical proportion.

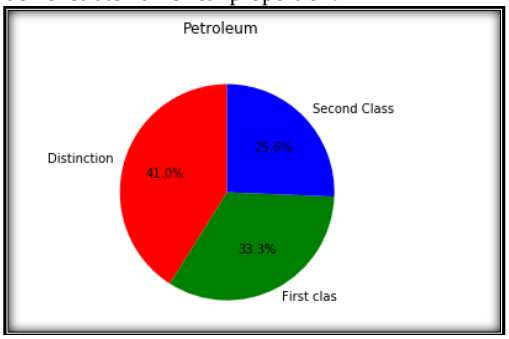


Figure.3.1 Pie chart of Class wise placement for Petroleum

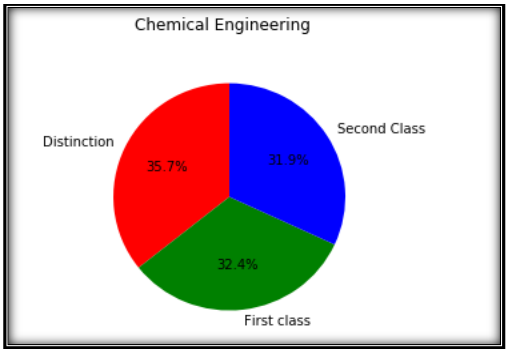


Figure.3.2 Pie chart of Class wise placement for Chemical Engineering

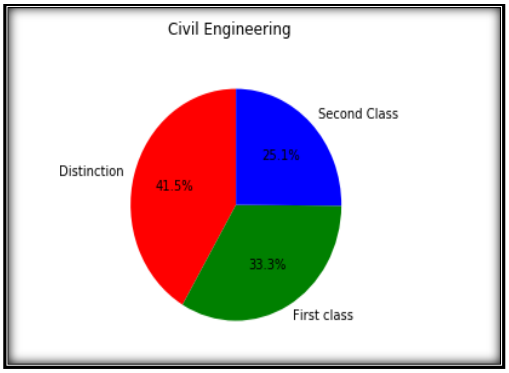


Figure.3.3 Pie chart of Class wise placement for Civil Engineering

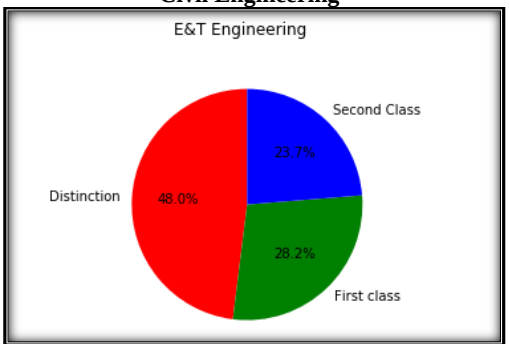


Figure.3.4 Pie chart of Class wise placement for E&TC Engineering

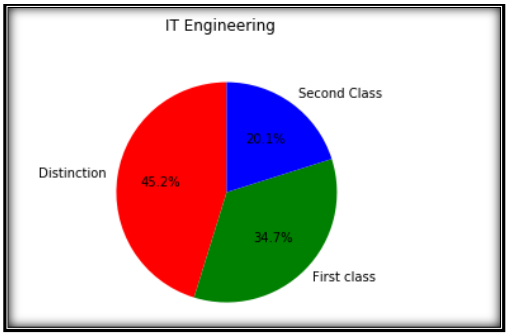


Figure.3.5 Pie chart of Class wise placement for IT Engineering

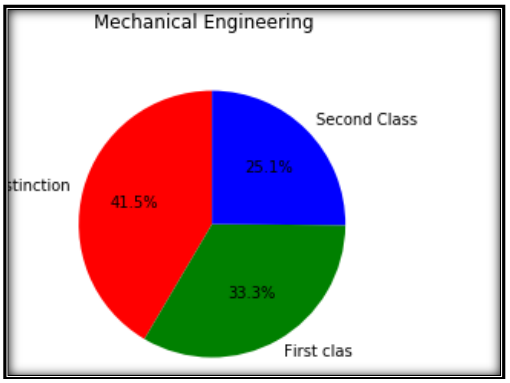


Figure.3.6 Pie chart of Class wise placement for Mechanical Engineering

The pie charts shown above give the information about students placed in campus interview based on their grade as per each stream. This gives an overview to students about stream and importance of aggregates according to the stream.

5. CONCLUSION

This paper examines application of Educational Data Mining (EDM). This paper elaborates the model to create awareness for students to create a better careeristic pathway for their future. Students with the help of their professors and placement team can make use of this model to get better placement opportunities and enhance their skillsets.

In Future this model can be compared with existing Machine Learning Algorithms like Linear Regression, Logistic Regression and Decision tree which will help us to understand the accuracy of

percentage of Machine Learning and Statistics. We will come to know what accurate percentage to rely on with the comparison of statistics and Machine Learning.

6. REFERENCES

[1] Dr. Mohd Maqsood Ali," ROLE OF DATA MINING IN EDUCATION SECTOR", IJCSMC, Vol. 2, Issue. 4, April 2013, pg.374 – 383

[2] Siddu P. Algur, Prashant Bhat and Nitin Kulkarni,"EDUCATIONAL DATA MININ-G: CLASSIFICATION TECHNIQUES FOR RECRUITMENT ANALYSIS" I.J. Modern Education and Computer Science, 2016, 2, 59-65

[3] K. Sreenivasa Rao, N. Swapna, P. Praveen Kumar "EDUCATIONAL DATA MINING FOR STUDENT PLACEMENT PREDICTION USING MACHINE LEARNING ALGORITHMS", International Journal of Engineering & Technology, 7 (1.2) (2018) 43-46

[4] V. Ramesh, P. Parkavi and P. Yasodha," PERFORMANCE ANALYSIS OF DATA MINING TECHNIQUES FOR PLACEMENT CHANCE PREDICTION", International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011.

[5] K. Sripath Roy 1, K. Roopkanth, V. Uday Teja, V. Bhavana, J. Priyanka," STUDENT CAREER PREDICTION USING ADVANCED MACHINE LEARNING TECHNIQUES", International Journal of Engineering & Technology, 7 (2.20) (2018) 26-29.

[6] Ajay Kumar Pal, "Classification Model of Prediction for Placement of Students", Published Online November 2013 in MECS (http://www.mecs-press.org/),DOI: 10.5815 /ijmecs.2013.11.07.

[7] Sudheep Elayidom, Dr. Suman Mary Idikkula and Joseph Alexander, "Applying Data Mining using Statistical Technique for Career Selection", International journal of Research Trends in Engineering, Vol. 1, No. 1, May 2009.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 394-396, 2019, ISSN:-2319–8656

Applications of Machine Learning for Prediction of Liver Disease

Khan Idris
Student (MSc Big Data Analytics)
MIT-WPU, Pune
Maharashtra,India

Sachin Bhoite
Assistant Professor
MIT-WPU, Pune
Maharashtra,India

Abstract: Patients in India for liver disease are continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. Therefore, building a model that will help doctors to predict whether a patient is likely to have liver diseases, at an early stage will be a great advantage. Diagnosis of liver disease at a preliminary stage is important for better treatment. We also compare different algorithms for the better accuracy.

Keywords: Indian Liver Patients, Machine Learning, Logistic regression, Support Vector Machine, Random Forest, AdaBoost, Bagging.

1. INTRODUCTION:

As there is growth in Liver Patients in India and it is estimated that till the year 2025 India may be the World Capital for Liver Diseases. We should have solution for this kind of problems and for this it is very important for doctors to identify the liver disease at an early stage. To identify liver disease, at an early stage we are building a machine learning model which will predict whether patient should be diagnosed or not at an early stage. We will be using different algorithms as well as ensemble methods. As, Liver disease can be diagnosed by analyzing the levels of enzymes in the blood. The objective of this model is to increase the survival rate of the Liver Patients by using the previous data about the levels of enzymes in their body. We have record of 583 patients from which 416 were the records of liver patient and 167 records of non liver patient.

2. RELATED WORK:

Ramana made a critical study on liver diseases diagnosis by evaluating some selected classification algorithms such as naïve Bayes classifier, C4.5, back propagation neural network, K-NN and support vector. The authors obtained 51.59% accuracy on Naïve Bayes classifier, 55.94% on C4.5 algorithm, 66.66% on back propagation neural network, 62.6% on KNN and 62.6% accuracy on support vector machine. The poor performance in the training and testing of the liver disorder dataset as resulted from an insufficient in the dataset [1]. We, have also gone through a research paper Diagnosis of Liver Disease Using Machine Learning Techniques by Joel Jacob1, Joseph Chakkalakal Mathew2, Johns Mathew3, Elizabeth Issac4. They have used FOUR classification algorithms Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbor (KNN) and artificial neural networks (ANN) have been considered for comparing their performance based on the liver patient data. Authors obtained 73.23% accuracy on Logistic Regression, 72.05% on k-NN, 75.04 accuracy on Support vector machine [2]. We have also gone through a paper Liver Patient Classification using Intelligence Techniques by Jankisharan Pahareeya, Rajan Vohra, Jagdish Makhijani, Sanjay Patsariya. In this paper Authors have used six intelligence techniques on the ILPD (Indian Liver Patient) Data Set. Throughout the study ten-fold cross validation is performed [3]. "Machine Learning Techniques on Liver Disease", in this paper authors have shown different types of techniques for disease prediction. Here algorithms Logistic Regression, SVM, Decision tree, Random Forest and ensemble techniques are used [4]. "Liver Classification Using Modified Rotation Forest", in this paper authors have gone through various classification algorithms to increase the accuracy and have done feature selection. Accuracy in this paper was 73.33% [5].

3. PROCESS IMPLEMENTATION:

The work flow process is firstly, we have to preprocess the data, then some data visualization part then we trained the model with different algorithms and selecting the algorithm with best output

.

3.1 Dataset

The Indian Liver Patient Dataset consists of 10 different attributes such as Age, Sex, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Phosphatase, Total Proteins Albumin, Albumin and Globulin Ratio, Dataset (result) of 583 patients. (416 records are of liver patients and 167 non liver patients). The patients were described as either 1 or 2 on the basis of liver disease. The feature Sex is converted to numeric value (0 and 1) in the data pre-processing step.

3.2 Data preprocessing

Data pre-processing is an essential step of solving every machine learning problem. It is said that 80% of the time of a Data Scientist is spend in data preprocessing. Most commonly used preprocessing techniques are very few like missing value imputation, encoding, scaling, etc. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. Therefore, in this column males are labeled as '1' and females are labeled as '0'.

The last column, Disease, is the label with '1' representing presence of disease and '2' representing absence of disease. This column is then relabeled as '1' for liver patients and '0' for the non liver patients. Total number of records is 583, with 416 liver patient records and 167 non-liver patient records. In the data visiualization of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with median values of the column.

3.3 Classification Techniques

a. Logistic Regression: Logistic regression is a type of a supervised machine learning algorithm. It makes a prediction that has binary outcome from the past data. Logistic regression usually returns result in very short time, hence it is preferred being used as a benchmarking model [4].

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = mx + c where the parameters are c and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X1 + ...PXp Where 0 is called the intercept [2]. As, it is seen that accuracy achieved by Logistic Regression was 73.23%, Now we are applying Adaboost to the Logistic Regression.

b. Support Vector machines: Support vector machines so called as SVM is a supervised learning algorithm which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process.

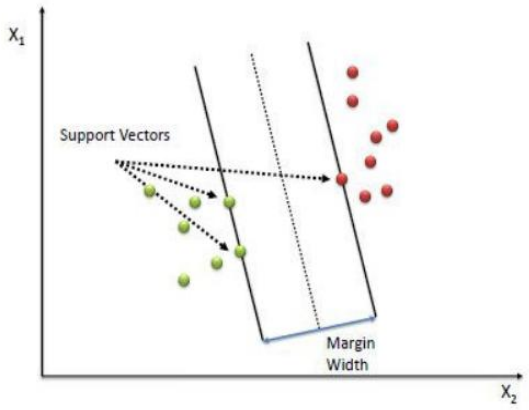


Fig.1 Support Vector Machine

c. Random Forest: Random Forest is a meta estimator that uses the number of decision tree to fit the various sub samples drawn from the original dataset, drawn data with replacement as per the requirements. Decision tree is non-parametric supervised learning. It is used for both classification and regression problem.

It is flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path between root and leaf represent classification rules. It creates a comprehensive analysis along with each branch and identifies decision nodes that need further analysis.

d.AdaBoost: AdaBoost is one of the first boosting algorithms to be adapted in solving practices. Adaboost helps you combine multiple "weak classifiers" into a single "strong classifier. The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. AdaBoost algorithms can be used for both classification and regression problem[6].

e.Bagging: Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

4. RESULTS AND EVALUATION:

The main objective was to predict whether a patient should be diagnosed or not at an early stage with algorithms such as SVM, Logistic Regression and random Forest. These algorithms were also used in previous studies. Now we have improves accuracy of these algorithms by using Bagging and AdaBoost.

Out[114]:

	Model	% Accuracy
0	Logistic Regression	73.504274
1	Support Vector Classfication	70.940171
2	Random Forest Classifiaction	66.666667
3	Adaboost Classifier Logistic	74.358974
4	Bagging Random Forest	72.649573

Fig. 2 Accuracies of Algorithms

As, we can see in the above figure increasing accuracies of the algorithms. We got accuracy 73.5% for Logistic Regression, then by applying Adaboost classifier the accuracy has been increased to 74.35%.

For Support Vector Machine we got 70.94%, and for Random Forest Classification 66.67% here we have got a considerable increase in accuracy by using Bagging that is the accuracy of 72.64.

5. CONCLUSION:

We have applied the machine Learning algorithms on the Indian Liver Patient dataset to predict the patients by the enzymes content in their at an early stage. We have used different machine learning classification algorithm as Logistic Regression, SVC, Random Forest and further we have applied bagging to Random Forest and AdaBoost to Logistic Regression. Logistic Regression is fast in processing and gave accuracy of 73.5%. Thus for increasing its accuracy we have used AdaBoost and got accuracy of 74.36%.

6. REFERENCES:

[1] Bendi Venkata Ramana1, Prof. M.Surendra Prasad Babu2 , Prof. N. B. Venkateswarlu3. “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis” International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.

[2] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac “Diagnosis of Liver Disease Using Machine Learning Techniques “by International Research Journal of Engineering and Technology (IRJET) 1,2,3 Dept. of Computer Science and Engineering, MACE, Kerala, India 4Assistant Professor, Dept. of Computer Science and Engineering, MACE, Kerala, India Volume: 05 Issue: 04 | Apr-2018.

[3] Jankisharan Pahareeya1, Rajan Vohra2, Jagdish Makhijani3 Sanjay Patsariya4 “Liver Patient Classification using Intelligence Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 2, February 2014 .

[4] V.V. Ramalingam1, A.Pandian2, R. Ragavendran3 “Machine Learning Techniques on Liver Disease - A Survey” 1,2,3Department of Computer Science and Engineering, SRMIST, Kattankulathur. International Journal of Engineering & Technology, 7 (4.19) (2018) 485-495.

[5] Bendi Venkata Ramana1 , Prof. M.Surendra Prasad Babu2 1 Associate Professor, “Liver Classification Using Modified Rotation Forest “Dept.of IT, AITAM, Tekkali, A.P. India. 2 Dept. of CS&SE, Andhra University, Visakhapatnam-530 003, A.P, India.

[6] https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe