VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

"Machine Learning algorithms for predicting the risks of chronic diseases"

Submitted in partial fulfilment for the award of degree

BACHELOR OF ENGINEERING IN
INFORMATION SCIENCE AND ENGINEERING

Submitted by:

NAME : GAGAN NAYAKA M N

USN : 1BI19CS055

innovation.creation

Conducted at

ICS

BANGALORE INSTITUTE OF TECHNOLOGY

Department of Computer Science

Accredited by NBA, New Delhi

Krishna Rajendra Rd, Parvathipuram, Basavangudi,Benagaluru
560004

BANGALORE INSTITUTE OF TECHNOLOGY

Department of Computer Science

Accredited by NBA, New Delhi

**Krishna Rajendra Rd, Parvathipuram, Basavangudi,Benagaluru
560004**

CERTIFICATE

This is to certify that the Internship titled **“Machine Learning algorithms for predicting the risks of chronic diseases”** carried out by **Mr. GAGAN NAYAKA M N ,** a bonafide student of Nitte Meenakshi Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering**, in COMPUTER SCIENCE AND ENGINEERING under Visvesvaraya Technological University, Belagavi, during the year 2022-2023. It is certified that all corrections/suggestions indicated have been incorporated in the report.

The project report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the course Internship / Professional Practice (18CSI85)

Signature of Guide **Signature of HOD** **Signature of Principal**

External Viva:

Name of the Examiner Signature with Date

1)__________________

____

2)__________________

____

DECLARATION

I, **Gagan Nayaka M N**, final year student of Computer Science, Bangalore Institute of Technology - 560 082, declare that the Internship has been successfully completed, in **ICS**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Computer Science, during the academic year 2022-2023.

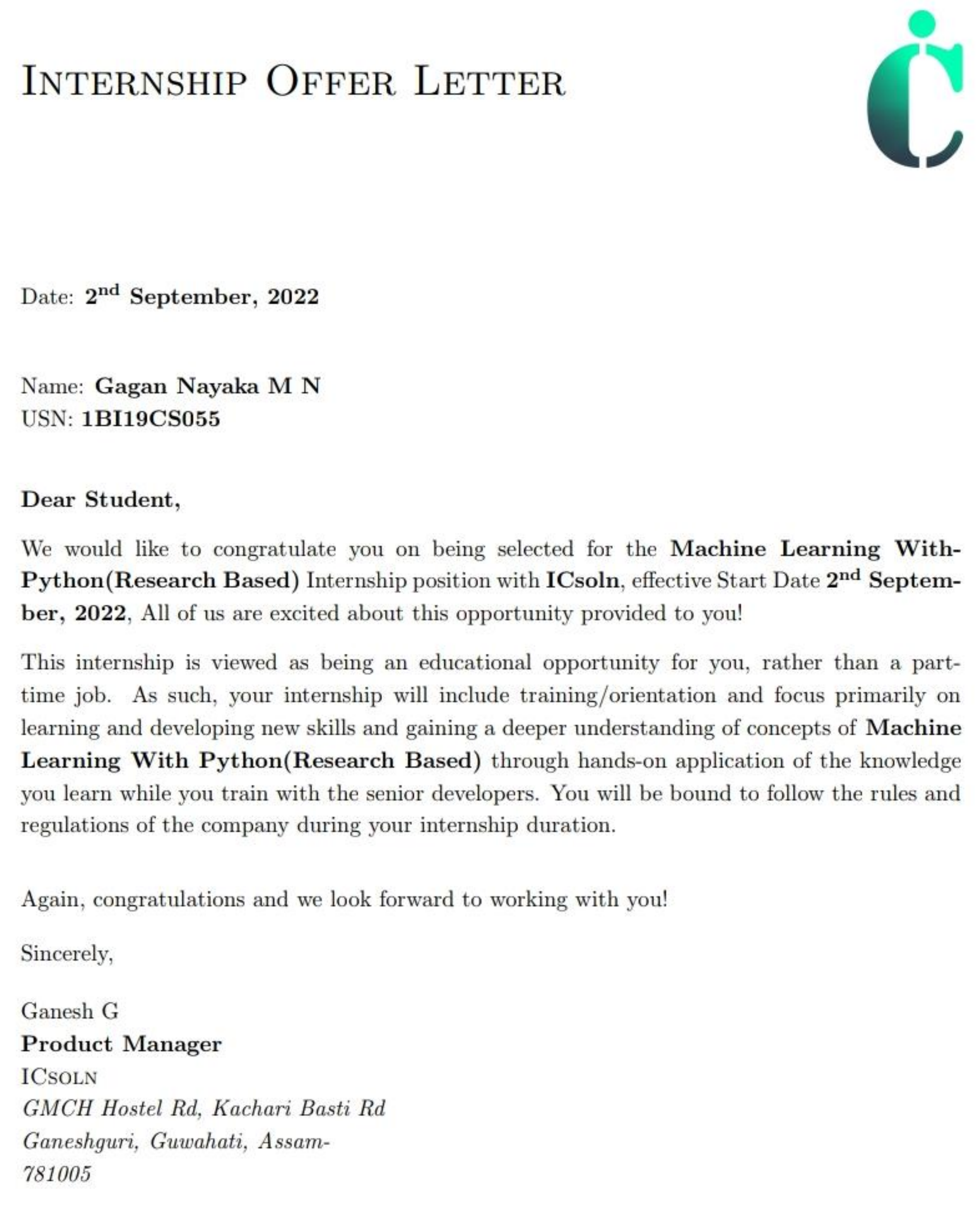
Date : 05.10.2022

Place : Bengaluru

USN : 1BI19CS055

NAME : Gagan Nayaka M N

OFFER LETTER



Internship Offer Letter

Date: **2$^{\text{nd}}$ September, 2022**

Name: **Gagan Nayaka M N**
USN: **1BI19CS055**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With-Python(Research Based)** Internship position with **ICsoln**, effective Start Date **2$^{\text{nd}}$ September, 2022**, All of us are excited about this opportunity provided to you!

This internship is viewed as being an educational opportunity for you, rather than a part-time job. As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts of **Machine Learning With Python(Research Based)** through hands-on application of the knowledge you learn while you train with the senior developers. You will be bound to follow the rules and regulations of the company during your internship duration.

Again, congratulations and we look forward to working with you!

Sincerely,

Ganesh G
Product Manager
ICsoln
GMCH Hostel Rd, Kachari Basti Rd
Ganeshguri, Guwahati, Assam-
781005

ACKNOWLEDGEMENT

This Internship is a result of accumulated guidance, direction and support of several important persons. We take this opportunity to express our gratitude to all who have helped us to complete the Internship.

We express our sincere thanks to Prof M U Aswath, our principal for providing usadequate facilities to undertake this Internship.

We would like to thank Dr J.Girija , our Head of Dept , for providing us an opportunity to carry out Internship and for his valuable guidance and support.

We would like to thank Software Services for guiding us during the period of internship.

We express our deep and profound gratitude to our guide,Prof Prashanth, Assistant/Associate Prof, for her keen interest and encouragement at every step in completing the Internship.

We would like to thank all the faculty members of our department for the support extended during the course of Internship.

We would like to thank the non-teaching members of our dept, for helping us during the Internship.

Last but not the least, we would like to thank our parents and friends without whose constant help, the completion of Internship would have not been possible.

GAGAN NAYAKA M N

ABSTRACT

Diabetes mellitus is one of the noxious disease which causes abnormalities of blood glucose due to the resistance of producing insulin hormone in the body. It affects various organs in the body such as the kidney, nerves, and eyes if it is not an early diagnosis.This total is expected to rise to 380 million within 20 years. Due to its importance, a design of a classifier for the detection of Diabetes disease with optimal cost and better performance is the need of the time.

The Pima Indian diabetic database at the UCI machine learning laboratory has become a standard for testing data mining algorithms to see their prediction accuracy in diabetes data classification. The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset.

The proposed model gives the closest results compared to clinical outcomes and also helps in the personalized diagnosis of patients. There are four machine learning algorithms these are Logistic Regression , Bernoullinve , Support Vector Machine (SVM), and Random Forest (RF) are used in the predictive analysis of early-stage diabetes.

Table of Contents

CHAPTER 1

COMPANY PROFILE

1. COMPANY PROFILE

A Brief History of ICS

ICS , was incorporated with a goal ”To provide high quality and optimal Technological Solutions to business requirements of our clients”. Every business is a different and has a unique business model and so are the technological requirements. They understand this and hence the solutions provided to these requirements are different as well. They focus on clients requirements and provide them with tailor made technological solutions.They also understand that Reach of their Product to its targeted market or the automation of the existing process into e-client and simple process are the key features that our clients desire from Technological Solution they are looking for and these are the features that we focus on while designing the solutions for their clients.

Sarvamoola Software Services. is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Sarvamoola Software Services. specialize in ERP, Connectivity, SEO Services, Conference Management, effective web promotion and tailor-made software products, designing solutions best suiting clients requirements.

ICS, strive to be the front runner in creativity and innovation in software development through their well-researched expertise and establish it as an out of the box software development company in Bangalore, India. As a software development company, they translate this software development expertise into value for their customers through their professional solutions.

They understand that the best desired output can be achieved only by understanding the clients demand better. Compsoft Technologies work with their clients and help them to defiine their exact solution requirement. Sometimes even they wonder that they have completely redefined their solution or new application requirement during the brainstorming session, and here they position themselves as an IT solutions consulting group comprising of high caliber consultants.

They believe that Technology when used properly can help any business to scale and achieve new heights of success. It helps Improve its efficiency, profitability, reliability; to put it in one sentence ” Technology helps you to Delight your Customers” and that is what we want to achieve.

CHAPTER 2

ABOUT THE COMPANY

2. ABOUT THE COMPANY

ICS is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, ICS specialize in ERP, Connectivity, SEO Services, Conference Management, effective web promotion and tailor-made software products, designing solutions best suiting clients requirements. The organization where they have a right mix of professionals as a stakeholders to help us serve our clients with best of our capability and with at par industry standards.They have young, enthusiastic, passionate and creative Professionals to develop technological innovations in the field of Mobile technologies, Web applications as well as Business and Enterprise solution. Motto of our organization is to “Collaborate with our clients to provide them with best Technological solution hence creating Good Present and Better Future for our client which will bring a cascading a positive effect in their business shape as well”. Providing a Complete suite of technical solutions is not just our tag line, it is Our Vision for Our Clients and for Us, We strive hard to achieve it.

Products of ICS

Android Apps

It is the process by which new applications are created for devices running the Android operating system. Applications are usually developed in Java (and/or Kotlin; or other such option) programming language using the Android software development kit (SDK), but other development environments are also available, some such as Kotlin support the exact same Android APIs (and bytecode), while others such as Go have restricted API access.

The Android software development kit includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on QEMU, documentation, sample code, and zutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows 7 or later. As of March 2015, the SDK is not available on Android itself, but softwaredevelopment is possible by using specialized Android applications.

Web Application

It is a client–server computer program in which the client (including the user interface and client- side logic) runs in a web browser. Common web applications include web mail, online retail sales, online auctions, wikis, instant messaging services and many other functions. web applications use web documents written in a standard format such as HTML and JavaScript,which are supported by a variety of web browsers. Web applications can be considered as a specifific variant of client–server software where the client software is downloaded to the client machine when visiting the relevant web page, using standard procedures such as HTTP. The Client web software updates may happen each time the web page is visited. During the session, the web browser interprets and displays the pages, and acts as the universal client for any web application. The use of web application frameworks can often reduce the number of errors in a program, both by making the code simpler, and by allowing one team to concentrate on the framework while another focuses on a specified use case. In applications which are exposed to constant hacking attempts on the Internet, security-related problems can be caused by errors in the program.

Frameworks can also promote the use of best practices such as GET after POST. There are some who view a web application as a two-tier architecture. This can be a “smart” client that performs all the work and queries a “dumb” server, or a “dumb” client that relies on a “smart” server. The client would handle the presentation tier, the server would have the database (storage tier), and the business logic (application tier) would be on one of them or on both. While this increases the scalability of the applications and separates the display and the database, it still doesn"t allow for true specialization of layers, so most applications will outgrow this model. An emerging strategy for application software companies is to provide web access to software previously distributed as local applications. Depending on the type of application, it may require the development of an entirely different browser-based interface, or merely adapting an existing application to use different presentation technology. These programs allow the user to pay a monthly or yearly fee for use of a software application without having to install it on a local hard drive. A company which follows this strategy is known as an application service provider (ASP), and ASPs are currently receiving much attention in the software industry.

Security breaches on these kinds of applications are a major concern because it can involve both enterprise information and private customer data. Protecting these assets is an important part of any web application and there are some key operational areas that must be included in the development process. This includes processes for authentication, authorization, asset handling, input, and logging and auditing. Building security into the applications from the beginning can be more effective and less disruptive in the long run.

Web design

It is encompasses many different skills and disciplines in the production and maintenance of websites. The different areas of web design include web graphic design; interface design; authoring, including standardized code and proprietary software; user experience design; and search engine optimization. The term web design is normally used to describe the design process relating to the front-end (client side) design of a website including writing mark up. Web design partially overlaps web engineering in the broader scope of web development. Web designers are expected to have an awareness of usability and if their role involves creating mark up then they are also expected to be up to date with web accessibility guidelines. Web design partially overlaps web engineering in the broader scope of web development.

Departments and services offered

ICS plays an essential role as an institute, the level of education, development of student's skills are based on their trainers. If you do not have a good mentor then you may lag in many things from others and that is why we at ICS gives you the facility of skilled employees so that you do not feel unsecured aboutthe academics. Personality development and academic status are some of those things which lie on mentor's hands. If you are trained well then you can do well in your future and knowing its importance of ICS always tries to give you the best.

They have a great team of skilled mentors who are always ready to direct their trainees in the best possible way they can and to ensure the skills of mentors we held many skill development programs as well so that each and every mentor can develop their own skills with the demands of the companies so that they can prepare a complete packaged trainee.

Services provided by ICS

- Core Java and Advanced Java
- Web services and development
- Dot Net Framework
- Python
- Selenium Testing
- Conference / Event Management Service
- Academic Project Guidance
- On The Job Training
- Software Training

CHAPTER 3

INTRODUCTION

3. INTRODUCTION

Introduction to ML

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.
Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience.The aim of this project is to develop a system which can perform detection of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques.

Problem Statement

To develop a machine learning model that detects the diabetes based on different parameters with a higher accuracy by combining the results of different machine learning methods.Also To provide medical professionals with a reliable prediction system to diagnose Diabetes and also detection system for non medical people(i.e citizens)

CHAPTER 4

SYSTEM ANALYSIS

4. SYSTEM ANALYSIS

1. Existing System

- There is no interactive tool for user to Detect diabetes.
- Using of different techniques the accuracy of detection is less.
- High false positives rates
- Early detection of diabetes which is the major concern, prioritizing the signs is not availabl

2. Proposed System

- To propose an effective technique for earlier detection of diabetes disease.
- To provide efficient model that detects the diabetes based on different parameters

3. Objective of the System

The aim of this project is to develop a system which can perform detection of diabetes for a patient with a higher accuracy by combining the results of machine learning techniques.

To provide medical professionals with a reliable prediction system to diagnose Diabetes and also detection system for non medical people(i.e citizens)

<u>CHAPTER 5</u>

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Hardware Requirement Specification

CPU	8th gen Intel Core i5 processor
RAM	12GB or higher
Disk Space	500 GB SSD or larger

Software Requirement Specification

Operating System	Windows 10 or 11
Platform	Python: 3.8.5

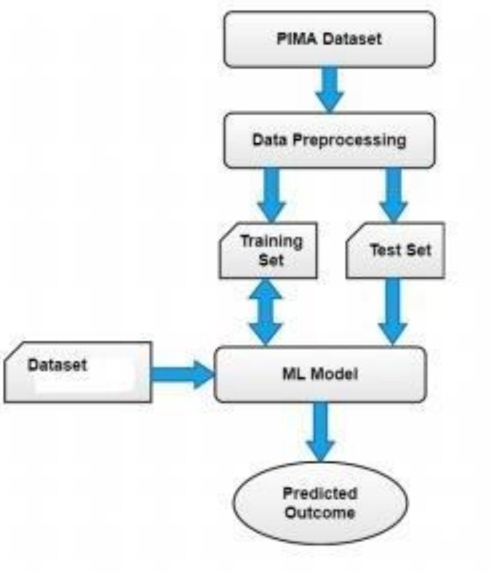
CHAPTER 6

DESIGN ANALYSIS

6. DESIGN & ANALYSIS

Architecture

The architecture is given in the below diagram. **Diabetes Detection** is a model that takes a set of data of both diabetics and non diabetic patients as an input and provides an output by classifiing the data using the SVM algorithm . This model trains the data with a certain percentages and stores the data as predicted training data. Among the 100% predicted training data it selects only a certain percent of badly predicted training data which will be trained again under the SVM model. The trained data is classified as diabeteic or non diabetic and gives the output.



Methodology

Methodology used for both diabetes detection and predictive analysis are as below
Classification techniques are widely used in pattern recognition or predictive analysis for classifying the data into different classes. Machine learning and artificial neural network technology are beneficial technologies that can do so due to the strength of their various classification algorithms supported by these technologies. These technologies are very frequently used in the medical field where predictive analysis is a challenging task; the causeof this is more imbalances and missing values in the data set. The procedures that are used in the building of a model contain several useful steps that are described below :

Data preprocessing: Machine learning algorithms are completely dependent on data because it is the most crucial aspect that makes model training possible. Initially when the dataset is collected from different sources is in the crude format so there may be a chance of many divergences, which the model may be unable to handle. So preprocessing is needed to remove all the divergences and prepare a clean data set. This included addressing missing values, calculating new features, and splitting data in the train-test set, data encoding means converting nonnumerical data into numerical data, normalizing data, etc. Another problem that occurs during the preprocessing phase is data imbalance; it means there exist more examples of one class than the other.

● Missing values: Missing values are those values such as the value for some attributes in the given sample is zero. To understand this let us take an attribute diastolic blood pressure containing zero value for a person is not possible . There are two approaches to solve the missing values problem

○ Deletion of record

○ Imputation method.

● Balance and unbalance dataset: In the classification area often data unbalanced problems occur and it is the problem of inequality in positive and negative predicted classes. If the number of positive samples is the same as the number of negative samples, then the dataset is said to be balanced otherwise it is unbalanced. The advantage of a balanced dataset is that the evaluation is easier to do since there is no bias

Data normalization: It is a very crucial aspect during the pre-processing phase. If you have a dataset, it may be the possibility of features of different units and scales [14]. In the given dataset some features are in low range scale and some are in high range scale, so for easy comparative analysis between them drawing them on the same scales and units is called normalization

Dataset Details

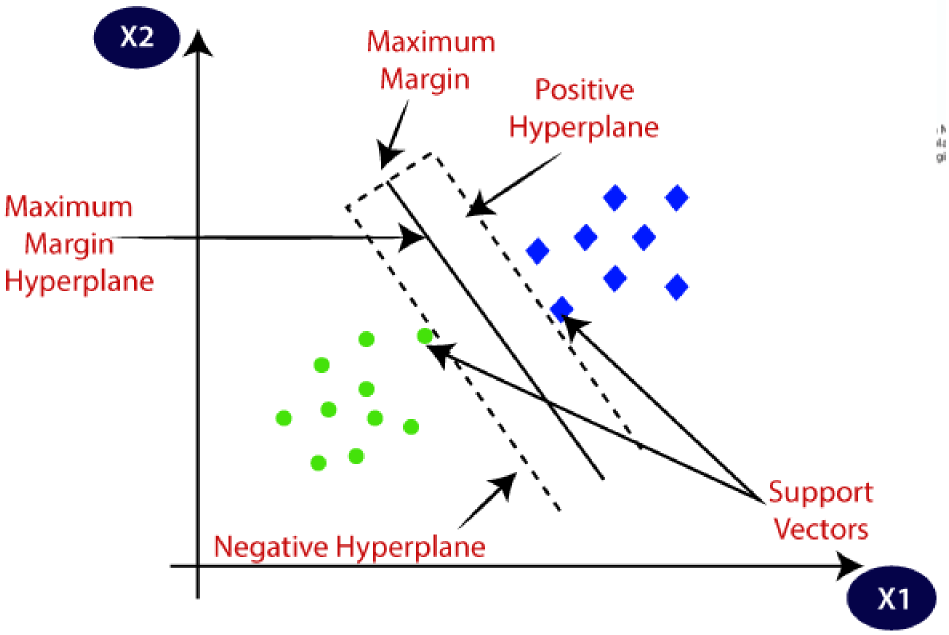
For the experimental analysis, a Pima Indian Diabetes Database (PIDD) is used taken from the University of California, Irvine (UCI) Repository that contains valuable features that are closely related to this disease [17]. This dataset comprises 768 records in which 268 positive predicted classes refer to diabetes patients and 500 negative predicted classes refer to non-diabetes is in the ratio of 34.9% and 65.1% of the whole dataset, respectively.There are eight numeric variables: (1) Number of times pregnant, (2) Plasma glucose concentration a 2h in an oral glucose tolerance test (3) Diastolic blood pressure (mm Hg) (4) Triceps skin fold thickness (mm) (5) 2-hour serum insulin (mu U/ml) (6) Body mass index (7) Diabetes pedigree function (8) Age (years). Although the dataset is labeled as there are no missing values, there were some liberally added zeros as missing values.

ML techniques used

For Detection of Diabetes we will be using :

SVM MODEL:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. SVM draws the data item in a higher dimension space. Suppose if you have ‘n’ features, it draws data items in n-dimensional space. SVM draws the hyperplane between dataset that best segregate the dataset into classes. The challenging task is the selection of the optimal hyperplane in the dimensional space and the right hyperplane is that plane which is on the highest margin between two classes. The points which are closer to the hyperplane are called the support vectors. The mapping of the objects is according to the specified boundaries of the hyperplane. The class of the new sample is based on a hyperplane that belongs to either one of the classes along the hyperplane .



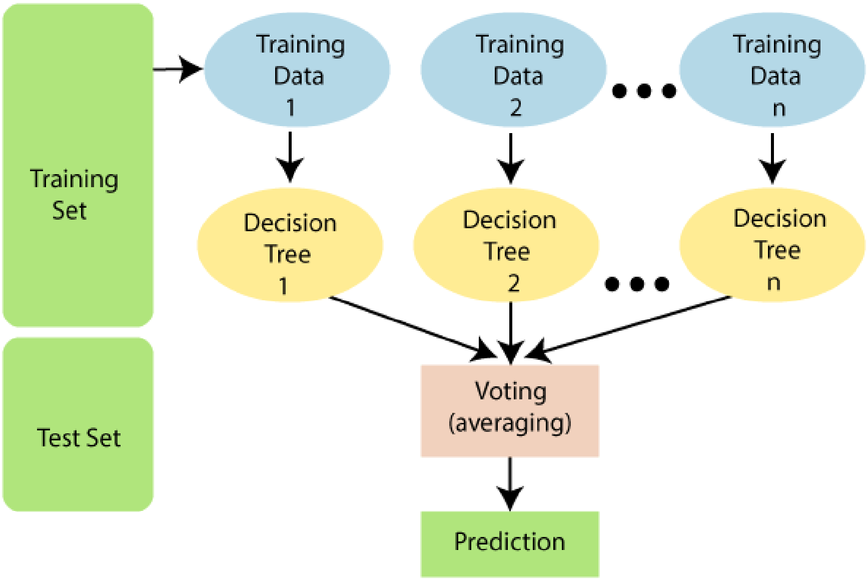
For early predictive analysis we have used following algorithms

Random Forest Algorithm:

RF is a very strong supervised learning algorithm, which is used in both cases classification as well as regression. It is an ensemble classifier that consists of a lot of decision tree and the prediction is based on the majority of votes collected from these trees [10]. So itgives better results in comparison to individual decision tree classifiers. It uses the concept of bagging technique to train each tree by generating the random sample features in the given sample. For generating the decision tree the commonly used algorithms are ID3 and CART.

Some useful steps that are used in RF are given below

- Initially load the training data that consists of ‘m’ features, which shows the behavior of the dataset.
- Randomly sample a subset of training (with replacement) called bagging such that select ‘n’ features randomly from ‘m’ features. x The ‘n’ training features are used in the modeling of ‘n’ decision tree.
- Gini index is used for the selection of splitting nodes (Best node) in the case of each decision tree.
- The above steps will go on for the modeling of ‘n’ number of the decision trees.
- The majority voted class is calculated in the count of collected votes of all trees in predicting the target class.
- Take the mode of all prediction in the case of classification and take the mean in the case of regression



R Naive Bayes Classifier Algorithm:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is mainly used in text classification that includes a high-dimensional training dataset.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with priorknowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where,
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

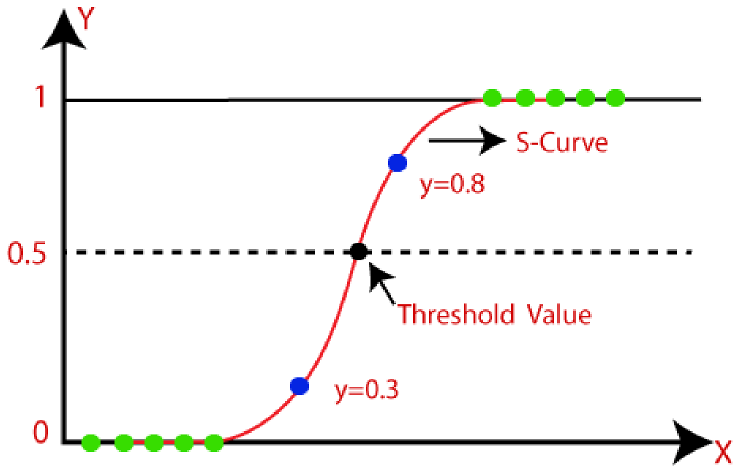
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

Logistic Regression :

Logistic Regression Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

- Logistic Regression is much similar to Linear Regression except that how they are used.
- Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data determine the most effective variables used for the classification.

The below image is showing the logistic function:



CHAPTER 7

IMPLEMENTATION

7. IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively.

The system can be implemented only after thorough testing is done and if it is found to work according to the specification. It involves careful planning, investigation of the current system and it constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning.

Two major tasks of preparing the implementation are education and training of the users and testing of the system. The more complex the system being implemented, the more involved will be the system analysis and design effort required just for implementation.

The implementation phase comprises of several activities. The required hardware and software acquisition is carried out. The system may require some software to be developed. For this, programs are written and tested. The user then changes over to his new fully tested system and the old system is discontinued.

TESTING

The testing phase is an important part of software development. It is the Information zed system will help in automate process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied. Software testing is carried out in three steps:

1. The first includes unit testing, where in each module is tested to provide its correctness, validity and also determine any missing operations and to verify whether theobjectives have been met. Errors are noted down and corrected immediately.

2. Unit testing is the important and major part of the project. So errors are rectified easily in particular module and program clarity is increased. In this project entire system is dividedinto several modules and is developed individually. So unit testing is conducted to individual modules.

3. The second step includes Integration testing. It need not be the case, the software whose modules when run individually and showing perfect results, will also show perfect results when run as a whole.

Implementation Details

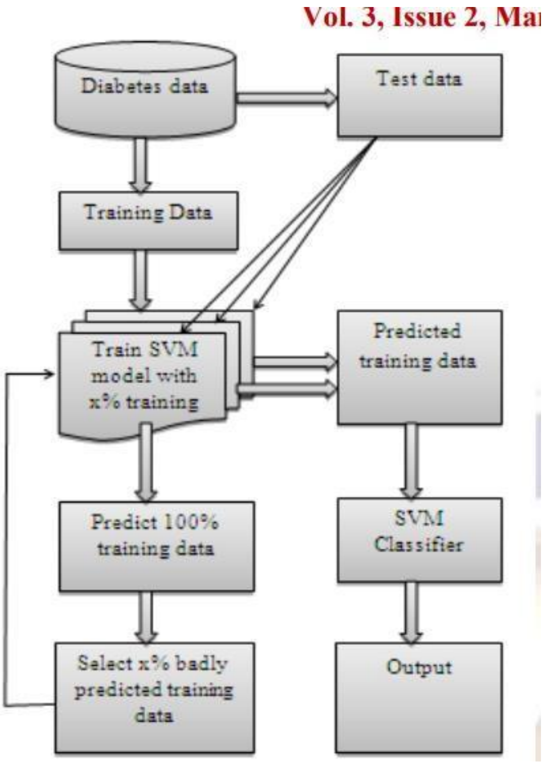
Languages/Tools/APIs used

Python is used for all of the modules. It has a good community supporting with various errors possible. Also, it has many ML libraries we can use. Our models come form python’s scikit-learn library and stats models library.

The following languages, tools and APIs were used in the project:

- Python
- Visual Studio/ Google Colab
- Python tools like matplotlib

Workflow Diagram


Vol. 3, Issue 2, Ma
Diabetes data
Test data
Training Data
Train SVM model with x% training
Predicted training data
Predict 100% training data
SVM Classifier
Select x% badly predicted training data
Output

Validation Methodology

To evaluate the robustness of the SVM models, a 10-fold cross-validation was performed in the training data set. The training data set is first partitioned into 10 equal-sized subsets. Each subset was used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the remaining nine datasets. This cross-validation process was repeated 10 times, and each subset served once as the test data set. Test data sets assess the performance of the models.

To analyze the performance of classification, the accuracy and AUC measures are adopted. Four cases are considered as the result of the classifier.

- TP (True Positive) : the number of examples correctly classified to that class.

- TN (True Negative): the number of examples correctly rejected from that class.

- FP (False Positive): the number of examples incorrectly rejected from that class.

- FN (False Negative): the number of examples incorrectly classified to that class. The classification experiments are conducted on the Diabetes dataset.

The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible values of the variables being classified

Table : Performance of SVM Classifier

Dataset	Accuracy	Sensitivity	Specification
Diabetes	78%	80%	76.8%

For predictive analysis the evaluation technique is :

K-fold cross-validation used to check the effectiveness and measure performance of the model. In this, the original dataset is prepared into a train-test set to validate the performance. Here K refers to the number of sections in which the whole data item is divided. To obtain the statistical reliable results experiments are conducted by several iterations. Suppose if the value of K is 10 the experiments will be conducted in 10 iterations. Out of K iteration for every valueof K one section is selected as a test set and the remaining K-1 sections are selected as a train set. The benefit of using this strategy is that each section gets an equal chance to become a test set. Take the mean of obtained results after K experiments which show the performance measures of the model.

The estimated mean error of the k tests is given by Equation 1.

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \qquad \mathbf{1}$$

Where E1 is the error obtained in each pass that occurs in the test dataset.

To measure the performance of various classifiers that are used to build a model, some important statistical metrics are calculated. The metrics are accuracy, sensitivity (recall), precision, specificity, and F-score. These metrics depend on classification labels such as true positive (TP), true negative (TN), false positive (FP), and false-negative (FN)

1. Accuracy: Divide the summation of TP and TN against the whole population.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \qquad (2)$$

2. Sensitivity / Recall: It measures the actual true positive rate and is calculated by using the formula given below.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

3. Specificity: It is the measure of the true negative rate and is given by the following formula.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (4)$$

4. Precision: It is defined by dividing the true positive against whole positive class values predicted. Mathematically it can be given by the following formula, given below

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (5)$$

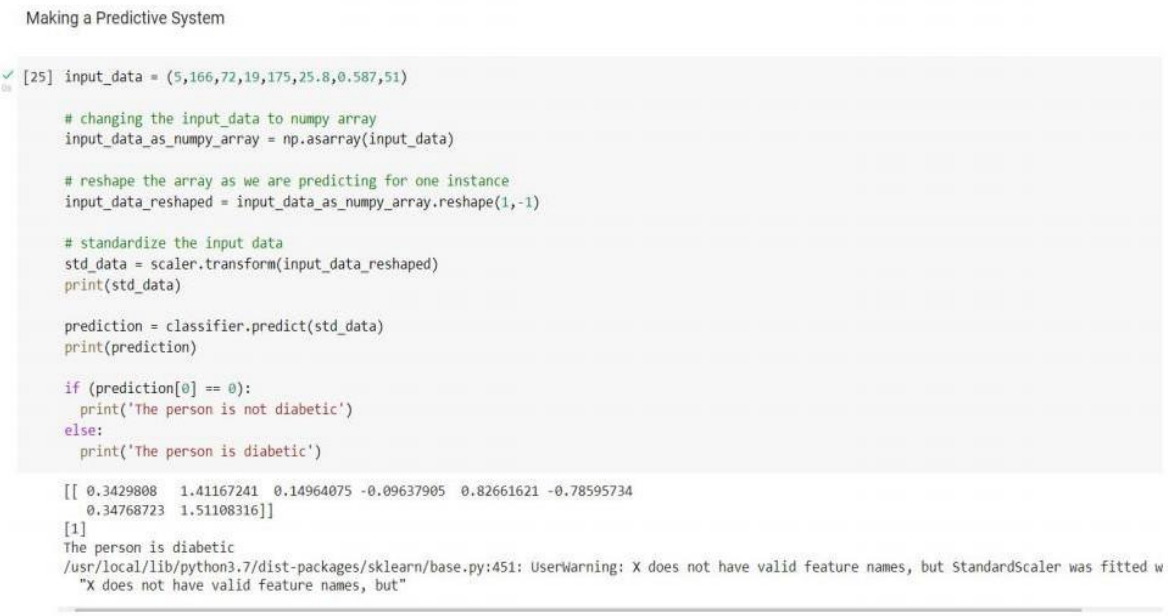
5. F-score: It is defined by dividing the true positive against whole positive class values predicted. Mathematically it can be given by the following formula, given below.

$$\text{F} - \text{score} = 2 \times \frac{\text{p} \times \text{r}}{\text{p} + \text{r}} \qquad (6)$$

CHAPTER 8

SNAPSHOTS

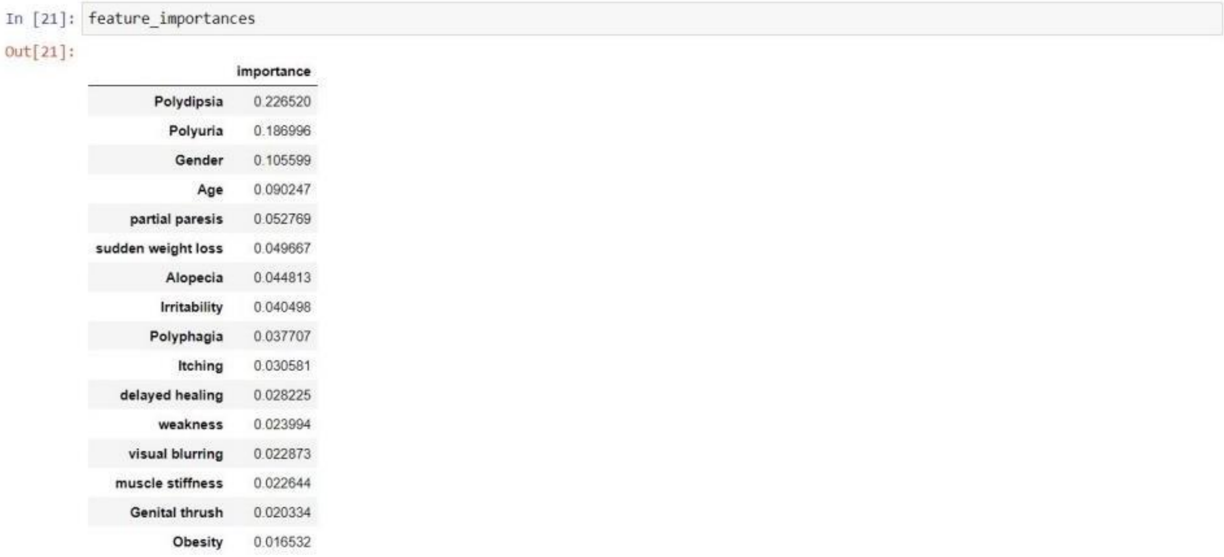
8. SNAPSHOTS



Making a Predictive System

```
[25] input_data = (5,166,72,19,175,25.8,0.587,51)

     # changing the input_data to numpy array
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the array as we are predicting for one instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     # standardize the input data
     std_data = scaler.transform(input_data_reshaped)
     print(std_data)

     prediction = classifier.predict(std_data)
     print(prediction)

     if (prediction[0] == 0):
       print('The person is not diabetic')
     else:
       print('The person is diabetic')
```

```
[[ 0.3429808   1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
   0.34768723  1.51108316]]
[1]
The person is diabetic
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but StandardScaler was fitted w
  "X does not have valid feature names, but"
```



```
In [21]: feature_importances
Out[21]:
```

	importance
Polydipsia	0.226520
Polyuria	0.186996
Gender	0.105599
Age	0.090247
partial paresis	0.052769
sudden weight loss	0.049667
Alopecia	0.044813
Irritability	0.040498
Polyphagia	0.037707
Itching	0.030581
delayed healing	0.028225
weakness	0.023994
visual blurring	0.022873
muscle stiffness	0.022644
Genital thrush	0.020334
Obesity	0.016532

CHAPTER 9

CONCLUSION

9. CONCLUSION

The package was designed in such a way that future modifications can be done easily. The following conclusions can be deduced from the development of the project:

- Automation of the entire system improves the efficiency
- It provides a friendly graphical user interface which proves to be better when compared to the existing system.
- It gives appropriate access to the authorized users depending on their permissions.
- It effectively overcomes the delay in communications.
- Updating of information becomes so easier
- System security, data security and reliability are the striking features.
- The System has adequate scope for modification in future if it is necessary.

10. REFERENCE

[1] Cortes, C., Vapnik, V., “Support-vector networks”, Machine Learning, 20(2),pp. 273-297, 1995.

[2] R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, pp. 1-5, 2017.

[3] Sneha, N., Gangil, T. ,”Analysis of diabetes mellitus for early prediction using optimal features selection,” in: Journal of Big Data 6, 13 (2019).

[4] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019.

[5] Birjais, R., Mourya, A.K., Chauhan, R., Kaur, H., “Prediction and diagnosis of future diabetes risk: a machine learning approach,” SN Appl. Sci. 1, 1112 (2019).

[6] M. Goyal, N. D. Reeves, S. Rajbhandari and M. H. Yap, "Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 4, pp. 1730-1741, July 2019.

[7] E. M. Aiello, C. Toffanin, M. Messori, C. Cobelli and L. Magni, "Postprandial Glucose Regulation via KNN Meal Classification in Type 1 Diabetes," in IEEE Control Systems Letters, vol. 3, no. 2, pp. 230-235, April 2019

[8] M. Goyal, N. D. Reeves, S. Rajbhandari and M. H. Yap, "Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 4, pp. 1730-1741, July 2019