Conference paper on “Prediction Of Stress And Diabetic Retinopathy Using Touch Screen Data ', is accepted for Thirteenth International Conference on Advances in Computer Science and Application – CSA 2024 will be held during Mar 29-30, 2024; in Bangalore, India.

Title of the project: Prediction Of Stress And Diabetic Retinopathy Using Touch Screen Data

Project Overview: The project aims to predict the presence of diabetes and stress using touch screen data by machine learning models. It employs data visualization, feature analysis, and multiple classifiers to build predictive models and assess their performance.

Abstract of the project: Touch screen data refers to the information collected from the interactions of users with touch-enabled devices, such as smartphones or tablets. Touch screen data can provide insights into the user’s behavior, mood, cognition, and health status. One of the new global concerns to public health is diabetes. Automated techniques are currently used to classify cases of Diabetes Retinopathy (DR). Therefore, it is essential to automatically identify human tension and relaxation using social media promptly. Early detection and effective management of stress are crucial in preventing its escalation into a severe condition. This paper presents a method to predict expressions of stress and Diabetic retinopathy by using ML techniques LR (logistic regression), RF ( random forest), and XG boost methods to detect the correlation between stress and diabetes while using touch screen.

Steps:

Data Loading and Exploration: Imports necessary libraries and loads the dataset ('projects.csv'). Prints 'success' after successful data loading. Displays basic information about the dataset and explores the target variables ('Diabetes', 'Stress').

Data Visualization: Uses various visualization techniques like histograms, pair plots, heatmaps, etc., to understand the distribution, relationships, and correlations among features and target variables. Utilizes seaborn and matplotlib for graphical representation.

Features Extraction and Model Building: Prepares the data by separating features (X) and target variables (Y) for both 'Diabetes' and 'Stress' predictions. Initiates and fits different classification models: Logistic Regression, Extra Trees Classifier, XGBoost, and Random Forest. Assesses the feature importances using Extra Trees Classifier to understand significant predictors.

Model Evaluation: Splits the data into training and testing sets using train\_test\_split. Evaluates the models' performances using accuracy\_score. Generates a results DataFrame showcasing the accuracy scores for each model.

Results Of the Project: Displays the accuracy scores achieved by each model for predicting 'Diabetes' and 'Stress'. Ranks the models based on their performance scores.

Conclusion: predicts stress and diabetes while using touch screens by applying different machine-learning techniques. In this work, machine learning techniques logistic regression, XG boost, and random forest algorithms were used to predict stress and diabetes using the touch screen. The result will give accurate results on how humans suffer from working prolonged hours with touch screens. The study systematically explores data, builds multiple predictive models, and assesses their accuracy for predicting diabetes and stress. The logistic regression predicts that people experience stress while using the touch screen for prolonged hours which leads to chronic diabate retinotherapy. The project systematically explores data, builds multiple predictive models, and assesses their accuracy for predicting diabetes and stress. The best-performing models are XGBoost and Random Forest for both 'Diabetes' and 'Stress' predictions, exhibiting accuracy scores of 0.86 and 0.76, respectively.

Description of Project: The growing use of digital gadgets has detrimental effects on both physical and mental well-being. Prolonged and consistent exposure may heighten stress levels, provoke anxiety, and disrupt sleep patterns, ultimately resulting in obesity, cardiovascular ailments, hypertension, impaired stress control, reduced levels of HDL cholesterol, and insulin resistance. Psychological consequences include suicide ideation, depressive symptoms, impaired sleep, and negativity impacted by content. Relying on digital devices may lead to heightened arousal, elevated levels of stress hormones, and impede mental vitality and growt This work used machine learning algorithms to forecast the occurrence of diabetes associated with stress in individuals who use touch screens. In this work, three models, namely RF (Random Forest), LR (Logistic Regression), and XG Boosting, were evaluated. The algorithms were trained using two datasets, which included measurements of glucose, insulin, blood pressure, BMI, stress, and age. The algorithms' accuracy was evaluated using performance measures. Users may retrieve the information to anticipate stress and diabetes by using touch-screen devices.

Stress disorders are a common issue among working IT professionals in the industry today. With changing lifestyle and work cultures, there is an increase in the risk of stress among the employees. Though many industries and corporates provide mental health related schemes and try to ease the workplace atmosphere, the issue is far from control. In this paper, we would like to apply machine learning techniques to analyze stress patterns in working adults and to narrow down the factors that strongly determine the stress levels. Towards this, data from the OSMI mental health survey 2017 responses of working professionals within the tech-industry was considered. Various Machine Learning techniques were applied to train our model after due data cleaning and preprocessing. The accuracy of the above models was obtained and studied comparatively. Boosting had the highest accuracy among the models implemented. By using Decision Trees, prominent features that influence stress were identified as gender, family history and availability of health benefits in the workplace. With these results, industries can now narrow down their approach to reduce stress and create a much comfortable workplace for their employees.

One of the new global concerns to public health is diabetes. The World Health Organisation (WHO) predicts that diabetes will rank as the seventh leading cause of death by 2030 (WHO, Diabetes, 2020). Depending on the ophthalmologist's experience, the diagnostic process can be difficult or time-consuming, especially in environments with limited resources. Automated techniques are currently used to classify cases of Diabetes Retinopathy (DR). This study aims to offer an automated DR detection system based on preprocessing, feature extraction, and classification procedures. Deep Convolutional Neural Networks (DCNN) and Machine Learning (ML) approaches are applied. About 62 million people in the Americas (422 million people worldwide) have diabetes, the majority living in low-and middle-income countries, and 284,049 deaths (1.5 million globally) are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. It is estimated that 62 million people in the Americas live with Diabetes Mellitus (DM) type2. This number has tripled in the Region since 1980 and it estimates that will reach the 109 million mark by 2040, according to the Diabetes Atlas (9th edition). Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. Poorly controlled diabetes increases the chances of these complications and premature mortality. In addition, people with diabetes are at higher risk of presenting cardiovascular diseases and tuberculosis, especially those with poor glycemic control.Globally, between 2000 and 2016, there was a 5% increase in premature mortality from diabetes. In the Americas, in 2019, diabetes was the sixth leading cause of death, with an estimated 244,084 deaths directly caused by diabetes. It is the second leading Disability Adjusted Life Years (DALY) cause, reflecting the limiting complications that people with diabetes suffer throughout their lives. Overweight/obesity and physical inactivity are major risk factors for diabetes type 2. The prevalence of overweight in the Americas was almost double that observed worldwide. Among adolescents in the Americas, 80.7% are insufficiently active. A healthy diet, regular physical activity, maintaining normal body weight and avoiding tobacco use are ways to prevent or delay the onset of type 2 diabetes. A healthy diet, regular physical activity, maintaining normal body weight and avoiding tobacco use are ways to prevent or delay the onset of type 2 diabetes. Diabetes can be treated and its consequences avoided or delayed with diet, physical activity, medication and regular screening and treatment for complications.

Type 1 diabetes (previously known as insulin-dependent, juvenile or childhood-onset) is characterized by deficient insulin production and requires daily administration of insulin. In 2017 there were 9 million people with type 1 diabetes; the majority of them live in high-income countries. Neither its cause nor the means to prevent it are known.Symptoms include excessive excretion of urine (polyuria), thirst (polydipsia), constant hunger, weight loss, vision changes, and fatigue. These symptoms may occur suddenly. Type 2 diabetes (formerly called non-insulin-dependent, or adult-onset) results from the body’s ineffective use of insulin. More than 95% of people with diabetes have type 2 diabetes. This type of diabetes is largely the result of excess body weight and physical inactivity.Symptoms may be similar to those of type 1 diabetes but are often less marked. As a result, the disease may be diagnosed several years after onset, after complications have already arisen. Until recently, this type of diabetes was seen only in adults but it is now also occurring increasingly frequently in children. Diabetes is a chronic condition that affects millions of people worldwide. It is characterized by high blood sugar levels due to the body's inability to produce or use insulin effectively. While there are various factors that contribute to the development of diabetes, including lifestyle choices and genetics, the Diabetes Pedigree Function is a tool that can help determine an individual's risk of developing the disease based on their family history.

Touch screens are used in numerous ways in an office setting, from swiping of IDs and inputting thumbprints to gain access, to actual office computers that require simple touch gestures for data entry and confirmation of different tasks. A touch screen is a computer display screen that is sensitive to human touch, allowing a user to interact with the computer by touching pictures or words on the screen. When using devices located in public spaces, we increasingly interact with them through touch screens. We operate in this manner devices such as ATMs, ticket machines, parking meters, parcel machines, and self-service checkouts. Although touch screens visually look similar, they are technologically different solutions, characterized by different ways of operation and factors such as sensitivity and precision, durability, or light permeability. Prolonged use of touch screens, especially in improper ergonomic positions, might lead to issues like eye strain, repetitive strain injuries, or musculoskeletal problems.

Prolonged usage of computer screens may result in many negative effects such as migraines, tiredness of the eyelids, reduced visual acuity, eye irritation, and discomfort caused by glare, insufficient lighting, or improper viewing angles. The prevalence of diabetes is increasing worldwide, with a projected rate of 7,079 cases per 100,000 inhabitants by the year 2030. This may be partially due to the widespread adoption of sedentary behaviors and the increasing dependence on cellular phones, television, and laptops. Deterioration of the retina's blood vessels leads to diabetic retinopathy, which manifests as fluid leakage, bleeding, and the development of new, more delicate blood vessels. Initially, it may result in minor visual impairments but may lead to complete loss of eyesight

Literature survey : Utilizing a deep learning architecture, An evaluation was carried out to assess the image quality of retinal imaging devices that are based on smartphones, including iExaminer, D-Eye, Peek Retina, and iNview. smartphone-based retinal imaging devices may serve as a substitute for direct ophthalmoscopes. However, it emphasizes that the field of vision plays a critical role in ensuring accurate automated identification. human behavior regarding smartphone touch interfaces to comprehend emotional states.

Propose methodology : The Study aims to predict the presence of diabetes and stress using machine learning models. It employs data visualization, feature analysis, and multiple classifiers to build predictive models and assess their performance. Prolonged exposure to screens, whether for leisure or work purposes, may lead to permanent injury in humans. Research indicates the negative effects of stress-related diabetes. Furthermore, with the progress of technology, people are becoming more immersed in the digital realm without their explicit agreement. Collecting and organizing more touch screen data from users with diabetes and DR, as well as their clinical and psychological assessments, to create a large and diverse dataset for training and testing the predictive models. Data is collected from freely available Kaggle and two datasets based on stress and diabetic retina therapy can use various sensors and devices to collect touch screen data from users, such as the frequency, duration, intensity, location, direction, speed, patterns, errors, corrections, usage, and context of touch interactions.

logistic regression is used for binary classification where we use the sigmoid function that takes input as independent variables and produces a probability value b/w 0 and 1.Sigmoid function is a mathematical function used to map the predicted values to probabilites.

Assumption of Logistic Regression:

\*Independent observations

\*No outliers in the dataset

\*Large sample size

\*Binary dependent variable

\*linear relationship b/w independent variable and dependent variable

How does Logistic Regression Work: the logistic regression model transform the linear regression functions continous value output into categorical value output using sigmoid function which maps the any real-valued set of independent variables input into a value b/w 0 and 1.this function is known as the logistic function.

TP=true positives,TN=true negatives,FP=false positives,FN=false negatives

->Accuracy: it provides the proportion of correctly classified Instance

->Accuracy=(TP+TN)/Total

->precision(p):Focuses on the accuracy of positive predictions

->p=TP/(TP+FP)

->Recall:It measures the correctly predicted positive instances among all actual positives instances.

->recall=TP/(TP+FN)

->F1 Score:harmonic mean of precision and recall

->f1 score=2\*(*precision\**recall)/(precision+recall)

logistic function to convert the output into a probability score

Linear and logistic have the difference of linear is measures continous outcome.And logistic function map predicted values b/w 0 and 1.

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.

bagging: where weak model are trained on different subsets of the trained data.each subset is sampled with replacement and prediction

boosting: in each model tries to correct the errors made by the previous models each models is trained on a modified version of the dataset the instances that were misused by the previous models are given more weight the final prediction is made by weighted voting.

Key features of Random forest:

\*high predictive accuracy

\*large dataset handling

\*handling missing value

\*parallelization for speed

\*addressing imbalanced data(adjusting class weight (or) resampling methods to ensure a balanced representation during training)

xgboost:It is the optimized distributed gradient boosting library designed for efficient and scalable training of ml models It is an ensemble learning method that combines the predicted of multiple weak models to produce a stronger prediction.It will handle large datasets in many machine-Learning tasks such as classification and Regression

It will have efficient handling of missing values.it possible to train models on large dataset in a resonable amount of time.

output: Diabetes->0,1

stress->1,2,3,4

Advantage: Performance, scalability,handling of missing value

disadvantage: computation complexity, Memory Requirement.

Knn: K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

\*data is collected from the various sensors and devices to collect touch screen data from users.

\*The person posts, stories, feeds of social media platform...etc data to help to predict the emotion of the people also.

\*The dataset is prepared according to the usage of the Touch Screen data

Compare to the other model accuracy xg boost is the best model because of ensemble learning and efficiently handles the missing values in the dataset.