**PROJECTREPORT**

***(Brain Stroke Prediction)*B. Tech IV year (Computer Science &Engineering)**

****

**Name – Sanskriti Mathur**

**Mentor – Mr. DebasishNandy**

**Celebal Program– COEDomain – Data Science**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude toMr. DebasishNandy for their invaluable guidance and support throughoutthecourse of this project. Their expertise, encouragement, andmentorship have been instrumental in shapingthedirectionandsuccess of this endeavor.

Mr. Debasish Nandy has been an exceptional mentor,consistently challenging me to think critically. Their dedicationto teaching and passion for the subject matter havebeenaconstant source of inspiration and motivation.

I would also like to thank Celebal Technologies for providingme this opportunity to gain experience whichhelpedmeinimproving my technical skills.

**ABSTRACTION**

The Brain Stroke Prediction Project is a data-driven initiative that aimstodevelop a predictive model to assess the risk of brain stroke basedonaset ofdemographic and health-related attributes. The project utilizes a dataset acquiredfrom Kaggle, containing valuable information such as gender, age, hypertension,heart disease, marital status, work type, residence type, average glucoselevel,BMI, and smoking status.

The data-set undergoes comprehensive data preprocessing, includinghandlingmissing values, encoding categorical variables, and feature scalingtoensuredataintegrity and model performance.

The core of the project lies in the implementation of a RandomForest Classifier,an ensemble learning algorithm that leverages multiple decisiontreestoenhanceprediction accuracy. The classifier is trained on a subset of the dataset andevaluated using accuracy as the primary metric, quantifying the proportionofcorrect predictions over the total number of instances. This evaluationstepisessential to gauge the model's effectiveness and its ability to generalizetounseendata.

The successful development and deployment of the Brain Stroke PredictionModel hold significant potential in the healthcare domain, enablinghealthcareprofessionals and individuals to proactively address stroke risk factors. Byproviding accurate and personalized predictions, the model can assist inmakinginformed decisions and promoting a healthier lifestyle. As a data-driventool,theproject exemplifies the power of machine learning and predictive analyticsinhealthcare research and preventive healthcare initiatives.

**INTRODUCTION**

The Brain Stroke Prediction Project is a cutting-edge application that aims torevolutionizethe field of healthcare by utilizing advanced machine learning techniques to predict thelikelihood of an individual experiencing a brain stroke. Brain strokes are a critical andlife threatening medical condition that requires immediate attention and intervention. Earlydetection and prevention are crucial in reducing the devastating impact of strokes onindividuals and society as a whole.

The project encompasses various steps, starting with data preprocessing, wheremissingvalues are handled, and categorical variables are encoded for model compatibility. Next,aRandom Forest Classifier, a powerful ensemble learning algorithm, is employedtobuildtheprediction model. The Random Forest Classifier's ability to handle complex interactionsandachieve high accuracy makes it an ideal choice for this task.

Furthermore, the model's performance is evaluated using appropriate metrics toassessitsreliability and generalization capabilities. This evaluation is crucial in ensuringthat themodel provides dependable predictions when deployed in real-world scenarios.

To make the prediction model accessible and user-friendly, we have integratedit intoawebapplication using Flask, a lightweight Python web framework. Users can input their demographic information and health-related attributes into an intuitive web form, andthemodel will swiftly generate a personalized stroke prediction.

The Brain Stroke Prediction Project has the potential to significantly impact thehealthcaresector by aiding in early stroke detection, prevention, and personalized care for at-riskindividuals. With the ability to predict stroke probabilities accurately, medical practitionerscan take proactive measures to improve patient outcomes and reduce the burdenof stroke related disabilities and mortality.

**IMPORTANCE OF BRAIN STROKE PREDICTION**

**Early Detection and Prevention**: Brain strokes are serious medical emergenciesthatcanlead to severe consequences if not detected and treated early. The Brain StrokePredictionModel plays a crucial role in identifying individuals at high risk of strokes basedontheirdemographic and health attributes. Early detection can lead to timely interventionandpreventive measures, potentially saving lives and minimizing long-termdisabilities.

**Public Health Impact:** Strokes are a leading cause of disability and deathworldwide.Bydeploying the Brain Stroke Prediction Model as a web application, it becomes accessibleto a wider audience. This includes individuals seeking information about their riskfactors,as well as healthcare professionals who can use the tool as an aid in their clinical assessments. The project contributes to the overall public health efforts toreducestrokeincidence and improve stroke management.

**Personalized Healthcare:** The model takes into account various factors suchasage,gender, lifestyle choices (e.g., smoking status), and health indicators (e.g., hypertension,heart disease). This allows for personalized predictions, empowering individualstomakeinformed decisions about their health and lifestyle choices. It also aids healthcareproviders in offering targeted interventions and recommendations for strokeprevention.

**Data-Driven Decision Making:** The Brain Stroke Prediction Model leveragesmachinelearning algorithms to analyze large datasets and identify patterns that maynot beapparent to human experts. By harnessing the power of data-driven decisionmaking,themodel can potentially improve the accuracy of stroke risk assessments andcomplementtraditional medical diagnostic methods.

**Reducing Healthcare Burden:** Brain strokes impose a significant burdenonhealthcaresystems, both in terms of treatment costs and resources. By enabling earlydetectionandprevention, the prediction model can help reduce the number of stroke cases, leadingtocost savings and alleviating the burden on healthcare facilities.

**OBJECTIVE**

The main objective of the Brain Stroke Prediction Project is to developapredictive model that can accurately assess the likelihood of an individual experiencing a brain stroke based on their demographic and health-relatedattributes. The key goals of the project include:

**Data Analysis:** Analyze and understand the provided dataset, whichcontainsinformation on gender, age, hypertension, heart disease, marital status, worktype,residence type, average glucose level, BMI, smoking status, andstrokeoccurrences.

**Data Preprocessing:** Perform necessary data preprocessing steps suchashandling missing values, encoding categorical variables, and scalingnumericfeatures to prepare the data for model training.

**Model Selection:** Choose an appropriate machine learning algorithmthat caneffectively predict the occurrence of a brain stroke based on the givenattributes.The Random Forest Classifier is selected for this task due to its abilitytohandlecomplex relationships and provide accurate predictions.

**Model Training:** Train the selected Random Forest Classifier onthepreprocessed data using a split of the dataset into training and testingsets.

**Model Evaluation:** Evaluate the performance of the trained model usingmetricslike accuracy, which measures the proportion of correct predictions.

**Model Deployment:** Integrate the trained model into a user-friendlywebapplication using Flask, allowing users to input their details andreceivepersonalized brain stroke predictions.

**DATA COLLECTION**

Data is the backbone of any data-driven prediction model. The dataset usedforbuilding the Brain Stroke Prediction Model was obtained fromKaggle.

Source : https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset

It contains the following features:

● gender: Gender of the individual (categorical: 'Male' or 'Female') ● age: Age of the individual (numeric)

● hypertension: Whether the individual has hypertension (binary: 0forNo,1forYes)

● heart\_disease: Whether the individual has a heart disease (binary: 0forNo,1for Yes)

● ever\_married: Marital status of the individual (categorical: 'Yes' or 'No')

● work\_type: Type of work of the individual (categorical: 'Private', 'Self-employed', 'Govt\_job', or 'children')

● Residence\_type: Type of residence of the individual (categorical: 'Urban' or'Rural')

● avg\_glucose\_level: Average glucose level in the individual's blood(numeric)● bmi: Body Mass Index of the individual (numeric)

● smoking\_status: Smoking status of the individual (categorical: 'neversmoked','formerly smoked', 'smokes', or 'Unknown')

● stroke: Whether the individual had a stroke (binary: 0 for No, 1for Yes)

**DATA PREPROCESSING**

Data preprocessing is a critical step in building theBrainStrokePrediction Model as it ensures that the dataset is properlycleaned,transformed, and ready for model training. Rawdataobtainedfromvarious sources may contain missing values, outliers,andinconsistencies. By handling missing values, encodingcategoricalvariables, and scaling numeric features, we preparethedataforsuccessful implementation of the RandomForest Classifier andaccurateprediction of brain stroke likelihood. Missing values areimputedusingvarious techniques, and outliers are identified and either correctedorremoved. Feature engineering is employed to derive newfeaturesthatcapture the complex interactions between air pollutantsandmeteorological variables.

Additionally, data normalization and scaling are appliedtobringallfeatures within a similar range, preventing any particular featurefromdominating the model training process.

**MODEL SELECTIONANDTRAINING**

The selection of an appropriate prediction model is crucialforachieving accurate and reliable results. Several machinelearningalgorithms, including Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks, areevaluatedfor their performance. Each model is trained usingaportionofthe dataset and evaluated on a separate validationset todetermine its predictive capabilities.

The Brain Stroke Prediction project involves predictingwhetheran individual is likely to experience a stroke basedonvariousdemographic and health-related attributes. For this predictiontask, we will use the Random Forest Classifier, a powerfulensemble learning algorithm known for its highaccuracyandrobustness.The Random Forest Classifier will be usedforthisprediction task due to its ability to handle non-linear relationships between features and the target variable. It alsoreduces overfitting by combining multiple decisiontrees.

**MODEL EVALUATION**

In the Brain Stroke Prediction Project, the model is evaluated usingvariousperformance metrics to assess its accuracy and effectiveness in predictingstrokeoutcomes. The primary evaluation metric used is accuracy, whichmeasurestheproportion of correctly predicted instances over the total number of predictions.However, it's essential to consider additional metrics for a comprehensiveassessment of the model's performance.

**1. Accuracy:**

Accuracy is calculated as the ratio of correctly predicted instancestothetotalnumber of instances in the test set. It provides an overall indicationof themodel'scorrectness in predicting stroke outcomes.

**2. Confusion Matrix:**

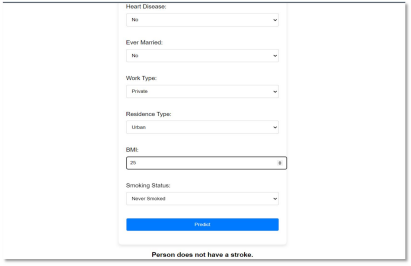
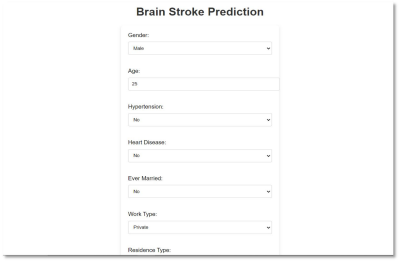
The confusion matrix is a valuable tool to assess the performanceofabinaryclassification model. It provides a detailed breakdown of the model'spredictions,distinguishing true positives (TP), true negatives (TN), false positives(FP),andfalse negatives (FN).

**3. Cross-Validation:**

To ensure the model's generalization capability, k-fold cross-validationcanbeperformed. It involves splitting the dataset into k subsets and usingeachsubsetasthe test set while training the model on the remaining data. Thisprocessisrepeated k times, and the average performance metrics are calculated.

**MODEL DEPLOYMENT**

The Model Deployment for the Brain Stroke Prediction project involvesconverting the trained machine learning model into a practical anduser-friendlyweb application. The goal is to make the predictive power of the model accessible to users without requiring them to interact directly withcodeordata.



**CONCLUSION**

Ultimately, the Brain Stroke PredictionModel signifiesaremarkable advancement in healthcaretechnology,providing a proactive approach to stroke prevention.Withits ability to assist in early diagnosis andriskassessment,healthcare professionals can utilize this tool toenhancepatient care and optimize resources. As thefieldofmedicalAI continues to evolve, the potential impact ofthisprediction model in reducing the burdenof stroke-relatedmorbidity and mortality cannot be understated. Assuch,itoffers a promising avenue for further researchandintegration into clinical practice to improvepublichealthoutcomes globally.

**REFERENCES**

1. **Kaggle Dataset: "Brain Stroke Prediction"** - The dataset usedforbuilding the Brain Stroke Prediction Model was obtainedfromKaggle.

2. **Python Programming Language -** Python was usedas theprimaryprogramming language for implementing the Brain StrokePredictionModel.

3. **Random Forest Classifier -** The RandomForest Classifierisanensemble learning algorithm used for building the predictionmodel.

4. **HTML and CSS -** HTML and CSS were used to createtheuserinterface of the web application. HTML tags were usedtostructuretheform elements and display the title and output, while CSSwasusedtostyle the web page and center the content on the screen.

5. **Scikit-learn Library -** The scikit-learn library in Pythonprovidedessential tools for data preprocessing, model building, andmodelevaluation. It offered a wide range of machine learningalgorithmsandevaluation metrics to support the development of the BrainStrokePrediction Model.

6. **Pandas Library -** Pandas was used for data manipulationandanalysis. It provided convenient data structures, such as DataFrames,which allowed for easy handling of the dataset and preparationformodel training.

7. **Numpy Library -** Numpy was used for numerical computationsandmathematical operations. It played a significant role inhandlingnumeric features like age, average glucose level, and BMI inthedataset.