**Group Members:**

Daniyal Shahid (20L-2095)

Anzla Aslam (20L-2109)

Aqsa Arshad (20L-1166)

**Project:**

**“Healthcare cost prediction”**

**Abstract:**

Clinical expenses are one of the most widely recognized repeating costs in an individual's life. It is general realized that an individual's way of life and various actual variables decide the illnesses or problems they might get, and that these circumstances decide clinical costs. As per a few examination, there are a few huge reasons that lead to more noteworthy consumptions. smoking, age, and BMI are factors in private clinical consideration. The objective of this study is to inspect and distinguish a connection between private clinical expenses and different qualities. Then, by producing direct relapse models and looking at them utilizing ANOVA, we utilize the huge qualities as indicators to gauge clinical consumptions.

In our examination, we found that smoking, age, and a higher BMI all have a critical association with higher clinical uses, showing that they are key supporters of the charges, and that the relapse can foresee the accuses of over 80% exactness. As per the World Wellbeing Association, individual clinical and medical care spending is becoming quicker than the worldwide economy This ascent in spending has been connected with various elements, the most unmistakable of which are smoking, maturing, and higher BMI. Utilizing protection information from different people with factors like smoking, age, HBA1C,heart issues, disease history, any transfer, number of kids and BMI, we desire to uncover a connection between clinical consumptions and different boundaries.

**Introduction:**

**Problem statement:**

In the domain of contemporary medical services, the raising intricacy of clinical benefits has prompted a basic requirement for exact medical care cost expectations. Our venture centers around fostering a prescient model to gauge individual medical care costs in view of a dataset of around 2000 people. This dataset incorporates factors, for example, a past filled with coronary failures, organ transfers, BMI, HBA1C levels, malignant growth history, number of significant medical procedures, and related charges. The test lies in making a vigorous model that obliges the variety of wellbeing boundaries as well as catches their nuanced cooperations. A definitive objective is to engage people for proactive monetary preparation and help medical care suppliers in upgrading asset designation, adding to more educated navigation. The intricacy of the issue emerges from precisely anticipating costs while unraveling multifaceted connections among assorted wellbeing factors. This undertaking tries to make a huge commitment to medical services cost forecast, upgrading the effectiveness of medical care frameworks and individual monetary preparation.

**Background:**

The worldwide medical services scene is going through a significant change described by progressions in clinical advancements, a maturing populace, and a rising predominance of ongoing illnesses. In the midst of these changes, the expense of medical services has risen considerably, introducing a critical test for people and medical services suppliers the same. It has become vital to comprehend and foresee medical care costs precisely to empower informed independent direction, asset advancement, and monetary preparation. As medical care costs keep on heightening, there is a developing requirement for information driven ways to deal with address this intricate issue. Information science and prescient demonstrating offer promising roads to unwind the unpredictable connections between different wellbeing boundaries and the monetary ramifications of ailments. In this specific circumstance, our undertaking tries to dive into the core of this test by bridling the force of information science procedures to foster a prescient model for assessing individual medical care costs.

The dataset, including information from very nearly 2000 individuals, fills in as a microcosm of the greater clinical consideration scene. It consolidates a scope of prosperity pointers, going from essential events, for instance, coronary disappointments and organ moves to more nuanced estimations like BMI and HBA1C levels. Understanding the components of these various factors and their total impact on clinical benefits charges is crucial for building a strong insightful model. The motivation driving this endeavor isn't simply to address the fast necessity for precise cost assumptions yet notwithstanding add to the greater chat on clinical consideration the board. By gaining pieces of information into the factors affecting clinical benefits costs, we mean to outfit individuals with a gadget for proactive money related readiness and help clinical benefits providers in improving resource conveyance. This, consequently, can provoke an additional plausible and capable clinical benefits structure that meets the creating necessities of the population.

In exploring the intricacies of the medical services scene, the mix of information science philosophies holds the possibility to upset how we approach medical services cost forecasts. Our task remains at the crossing point of medical care, information science, and direction, with the goal to cultivate a more profound comprehension of the expense elements in medical care and prepare for more educated and powerful medical care the board methodologies.

**literature review:**

The writing on medical care cost forecast is huge and multi-layered, mirroring the assorted endeavors to foster precise models that can disentangle the complexities of medical care consumptions. An exhaustive survey of existing examinations gives important bits of knowledge into the systems, difficulties, and progressions in this basic space.

* Traditional Approaches : Traditional linear regression models have been broadly utilized in medical services cost forecast, giving an essential comprehension of the connections between individual wellbeing boundaries and related costs. Traditional linear regression relapse frequently center around unambiguous infections or wellbeing measurements, endeavoring to lay out direct connections. While these models offer effortlessness and interpretability, they might battle to catch the nonlinear and complex nature of medical services cost elements.
* Decision Trees and Ensemble Methods: Decision trees and ensemble methods, such as Random Forests, have acquired prevalence for their capacity to deal with nonlinearity and catch complex connections between factors. These models, described by their tree-like designs, have been applied to anticipate medical services costs by considering different wellbeing factors all the while. Group techniques, specifically, mean to improve prescient exactness by collecting the results of numerous choice trees. While viable, these models might confront difficulties in taking care of enormous and high-layered datasets.
* Deep Learning Approaches: The advent of deep learning has introduced another time for medical services cost expectation. Deep neural networks with their ability to automatically learn hierarchical features from data, have shown guarantee in catching unpredictable connections that might evade customary models. Concentrates on integrating profound advancing frequently influence neural network architectures to simultaneously examine assorted wellbeing boundaries. The adaptability and flexibility of profound learning models make them appropriate for medical services cost expectation errands, particularly while managing huge datasets and complex connections.
* Machine Learning Techniques: Machine learning, beyond deep learning, continues to play a pivotal role in healthcare cost prediction. Support Vector Machines , Gradient Boosting and other regression based algorithms have been explored in various studies. These models characterized by their versatility and ability to handle diverse data types, contribute valuable insights into the factors influencing healthcare expenditures.
* Holistic Approaches and Challenges: While existing examinations have taken critical steps, a typical pattern arises - the requirement for additional comprehensive methodologies. Many examinations center around unambiguous infections or segregated wellbeing boundaries, possibly ignoring the interconnected idea of medical services elements. Challenges incorporate the requirement for huge, various datasets that enough address the populace, resolving issues of information quality and culmination, and guaranteeing model interpretability for viable execution in in healthcare settings.
* Integration of Data Sources: Recent literature emphasizes the integration of diverse data sources, including electronic wellbeing records, hereditary data, and patient way of life information, to improve the granularity and prescient force of models. By joining these rich datasets, scientists expect to foster far reaching models that exemplify the complex idea of medical care.
* Emerging Trends: Emerging trends in literature include the exploration of artificial intelligence models, moral contemplations in medical care cost expectation, and the joining of constant information streams for dynamic expectations. As the field develops, analysts progressively perceive the significance of foreseeing costs precisely as well as guaranteeing the interpretability and moral arrangement of these prescient models in genuine medical care situations.

In summary, the literature on healthcare cost prediction reflects a dynamic landscape where traditional and advanced computational techniques coexist. The evolution from linear regression to sophisticated deep learning models signifies a paradigm shift in understanding and addressing the complexities inherent in healthcare cost dynamics. Notwithstanding, challenges endure, accentuating the continuous requirement for research that coordinates assorted information sources, thinks about all-encompassing methodologies, and addresses moral contemplations for the functional execution of prescient models in medical services settings. The examination introduced in this paper adds to this advancing story by investigating relapse based profound learning and AI calculations for medical care cost expectation.

**Critical Analysis:**

* Lets see, most importantly, first paper having title "Predictive Modeling of Healthcare Costs " .This paper broadly surveys different prescient displaying strategies, including direct relapse, choice trees, and group techniques. It surveys their materialness to medical services cost expectation and features the significance of consolidating assorted wellbeing boundaries in the demonstrating system. The review uncovers that troupe techniques, like Irregular Timberlands, will generally beat individual calculations in anticipating medical care costs. Furthermore, it highlights the requirement for include designing to improve the model's exactness. The paper, be that as it may, may need particularity in tending to the extraordinary difficulties related with wellbeing information, like managing missing qualities or slanted disseminations. Executing the discoveries might require extra contemplations well defined for medical services datasets.
* The following one with title" Machine Learning Approaches to Health Cost Prediction " Focused on AI, this examination investigates the use of calculations in foreseeing medical care costs. It dissects the job of highlights like BMI, clinical history, and medical procedures in foreseeing costs, giving a nuanced comprehension of their effect. The review distinguishes BMI as an especially powerful component and underscores the significance of catching non-straight connections in wellbeing cost forecast. It additionally recommends that more mind boggling models might be important to catch the intrinsic intricacy of medical care information. The paper may not widely address interpretability gives that can emerge with complex models. Finding some kind of harmony between model intricacy and interpretability is vital for commonsense execution in medical services settings.
* The following one having title " Impact of Chronic Diseases on Healthcare Costs" ,This concentrate fastidiously inspects the connection between's persistent infections and medical care costs. It dissects the particular variables inside ongoing circumstances that fundamentally impact costs, giving a granular comprehension of their effect. The examination recognizes an unmistakable relationship between the seriousness of constant illnesses and raising medical care costs. It features the significance of thinking about the drawn out monetary ramifications of constant circumstances in prescient models. The paper may not unequivocally address the difficulties of managing imbalanced datasets, where ongoing circumstances may be underrepresented. Tending to such lopsided characteristics is pivotal for a more exact portrayal of certifiable situations.
* The next one with tittle ” Ethical Considerations in Healthcare Cost Prediction” ,Focused on the moral components of medical care cost expectation, this survey investigates issues like information security, predisposition, and straightforwardness. It gives a basic focal point on the moral ramifications of prescient displaying in medical services. The paper highlights the significance of keeping up with patient protection, limiting predispositions in prescient models, and guaranteeing straightforwardness in model turn of events and arrangement. While featuring moral contemplations, the paper may not give substantial arrangements or rules to relieving these moral difficulties. Carrying out moral norms in prescient demonstrating requires a nuanced comprehension of medical services explicit moral situations.
* The examination paper “Integration of Behavioral Factors in Healthcare Cost Prediction" digs into the consolidation of social elements in medical services cost forecast models. It investigates how way of life decisions, propensities, and patient ways of behaving add to varieties in medical services consumptions. The review stresses the meaning of catching these elements for a more comprehensive comprehension of wellbeing cost elements. It proposes that prescient models consolidating social information can give more customized and exact quotes. Be that as it may, the paper may not widely examine the difficulties related with getting and incorporating social information into prescient models, for example, guaranteeing information dependability and addressing potential predispositions connected with self-detailed data.
* The last however not the least with title " Data-Driven Insights into Healthcare Expenditure " ,This new expansion to the writing investigates certifiable contextual analyses where prescient demonstrating has effectively improved medical services asset assignment. It gives commonsense bits of knowledge into the use of information driven approaches in medical services consumption. The paper stresses the effect of prescient displaying on asset improvement, exhibiting cases where precise expense expectations have prompted more productive portion of medical care assets. The review may not completely talk about possible difficulties in sending prescient models in different medical services settings. Carrying out fruitful procedures might require adjusting to the special qualities of various medical services conditions.

In conclusion, The explored papers all in all contribute important bits of knowledge to the field of medical services cost expectation, each tending to explicit angles and difficulties related with prescient displaying. We saw the procedure, key discoveries and the limit of every one of the model so we can beat the issues looked by these models by making the new exact model for cost assessment of wellbeing illnesses of a person with more precision. The writing underscores the variety of variables affecting medical care costs, going from AI methods and ongoing infection contemplations to moral ramifications and transient elements. the writing survey uncovers the diverse idea of medical care cost forecast, with each paper adding to a more far reaching comprehension of the difficulties and potential open doors in this developing field. As the exploration advances, tending to these difficulties and consolidating different points of view will be essential for the improvement of powerful and morally sound prescient models in medical care cost assessment. The mix of social elements, thought of transient elements, and a persistent obligation to moral norms will improve the precision and materialness of prescient models in genuine medical services settings.

**Methodology:**

There are 5 modules in this project that have detailed explanation below:

**1) Data Collection:**

The "Healthcare Cost Prediction Dataset" is a collection of data related to healthcare costs for various individuals. It contains information about the demographics, habits, diseases and other attributes of patients, as well as the charges they incurred for healthcare services. This dataset is valuable for building predictive models to estimate healthcare costs and gain insights into the factors influencing these costs. We get three csv files from the source given above first we concatenate these 3 files and then we applied all the operations mentioned in the assignment in order to make the dataset ready for analysis. We cleaned the data set in which we have charges column that is directly related to other columns which are basically diseases so as our model is healthcare cost prediction so we will train our model in which based on this data set our model will successful predict the cost that an individual will have to pay for his health treatment. want to investigate the relationships between different attributes and healthcare costs to identify key factors influencing expenses. We had utilized a dataset from Kaggle Site for making our expectation model. This informational index incorporates nine ascribes and the informational index has splitted into two sections, preparing information and testing information. For preparing the model, 80% of all out information is utilized and the rest for testing. To fabricate an indicator model of clinical protection cost the preparation dataset is applied and to assess the relapse model, test set is utilized.

The dataset likely includes a variety of attributes that can be used for healthcare cost prediction. Customer ID: The ID of the individual.

BMI: Body Mass Index, a measure of body fat based on height and weight.

HBA1C: It is a blood test that measures the average blood sugar levels over the past 2- 3 months and is used to assess long-term blood glucose control, particularly in diabetes management.

Heart Issues: A binary indicator of whether the individual has a heart\_issues (yes/no).

Any Transplants: A binary indicator of whether the individual has any transplant in the past.

Cancer history: A binary indicator of whether the individual has cancer of any type in past.

Number Of Major Surgeries: Numbers of major surgeries of individual.

Charges: The actual healthcare charges incurred by the individual. This is the target variable for prediction.

Smoker: A binary indicator of whether the individual is a smoker (yes/no).(float) Our model will predict the charges against other features which are diseases and other records that are useful in predicting the charges against the information he gave to model

**2)Data pre-processing:** The data that will be utilized to answer the problem is one of the most significant aspects of machine learning difficulties. Data preparation accounts for around sixty to seventy percent of the overall time spent on a typical machine learning project. In order to get successful outcomes, it is critical to have the proper data for the situation at hand. In general, data preparation consists of selecting characteristics and pre-processing those features. As a result, after selecting features from a vast quantity of data, the following step is to pre-process those features. Because the data is useless in its raw form. The goal of pre-processing in this case is to make features appropriate for the machine learning model we'll use. If the characteristics are set up correctly, the model can produce better results.

**Data Cleaning:** Many contaminants can be found in raw data. These contaminants can have an impact on the ultimate result, especially when it comes to machine learning difficulties. As a result, after the data has been incorporated, it must be cleansed. Impurities such as incorrect entries, irrelevant data, and inconsistent data are detected during data cleaning. Eliminating these contaminants from records Data cleansing may be accomplished using a variety of methods. Using automated tools, personal interaction, and building scripts to programmatically clean the data according to our needs are some of the data cleaning strategies. To ensure the robustness of the predictive model, the dataset undergoes a rigorous preprocessing phase:

* **Handling Missing Values**: Robust imputation techniques are applied to address missing values, preserving the integrity of the dataset. Like we have missing values in heart issues , no of children’s and some more columns. So in handle these missing values we applied missing values techniques like forward and backward fill , regression imputation and multiple imputation . these techniques are used based on the most suitable technique for the column like as smoking is irrelevant to others attribute like heart issues and other diseases so for this we used multiple imputation. Same in the case for forward and backward fill this is used for the attributes like heart issues where missing values are completely random , regression imputation where there is randomness in the missing values .
* **One-Hot Encoding:** Machine learning algorithms often work with numerical inputs, and many models require numeric representations of features. One-hot encoding converts categorical variables into binary vectors, where each category is represented by a binary bit. This binary representation allows the model to understand and process categorical information. Categorical variables, such as no of surgeries, month, city tier and name of individuals are encoded using one-hot encoding to convert them into a format suitable for machine learning models. Like here we convert our data to numeric from categorical.
* **Handling Duplicates:** There are many entries in the data which are repeated entries like having multiple customer id entries and caused a redundancy in the data so we removed the redundancy in order to handle the duplicate problem for accurate model training.

**Transformation Data**: After cleaning data and integration data, the following is step to convert the Clean data that has been incorporated into the system's format. In most cases, data conversion entails transforming the target data into the format needed by the source data.

Normalization and Scaling: Numeric features, including health metrics, are normalized to a standard scale to prevent any particular variable from dominating the model due to its magnitude. So after cleaning the data we apply transformation technique name as normalization in order to convert it in system format for integration.

**Dimensionality Reduction:**

Working with too many variables is expensive and may lead to unexpected result. It is referred as the Curse of Dimensionality. Look through your dataset and use an appropriate Dimensionality Reduction Algorithm to cut down the variables that are not useful for data modeling step. There are a lot of irrelevant and useless columns that are not for our model like city tier, month, date , hospital tier , city tier . These are the columns which are totally useless because we need only disease entries , smoking and charges against these , all other than these are irrelevant . so by Dimensionality Reduction we remove the unnecessary columns.

**3)Model Selection:**

Our methodology incorporates a hybrid approach, leveraging the strengths of both deep learning and traditional machine learning algorithms:

* **Deep Learning Model:** A neural network architecture is designed to capture complex relationships within the dataset. The deep learning model consists of multiple layers, incorporating appropriate activation functions, regularization techniques, and optimization algorithms to effectively learn intricate patterns.
* **Machine Learning Model**: Concurrently, traditional machine learning regression models, such as linear regression, decision trees, and ensemble methods, are implemented. These models serve as benchmarks for comparison against the deep learning counterpart.

**4)Model Training:**

The dataset is partitioned into training and testing sets to facilitate model training and evaluation:

* **Training Process**: The deep learning model and machine learning models are trained on the training dataset, iterating through multiple epochs to optimize weights and biases.
* **Hyperparameter Tuning:** Parameters for both deep learning and machine learning models are tuned to achieve optimal performance, striking a balance between underfitting and overfitting.
* **Cross-Validation:** K-fold cross-validation is employed to enhance the robustness of the models, ensuring that they generalize well to unseen data.

**implementation:**

**results:**