Section A-Research paper
ISSN 2063-5346

The Human Freedom Index: Exploratory Data Analysis and Target Prediction

Sahana Ashok (Corresponding Author)
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology
Puducherry, India
lochanaa24@gmail.com
ORCID: 0009-0000-8576-7080
LinkedIn: https://www.linkedin.com/in/sahana-a/

R. Raj Bharath
(Assistant Professor) Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology
Puducherry, India
rajbharathraj@gmail.com

***Abstract*— Individual dignity and rights are recognized by the extent of human freedom; an essentially valued social idea that allows and empowers people to follow their choice. The Human Freedom Index is an annual assessment that quantifies the level of human freedom by presenting a broad measure among various aspects including areas like Rule of law, Security, Religion, Expression, Regulation, etc. Given these measures for each country, the central objective is to provide an overall view of the extent of freedom worldwide, its projections, and changes. Based on these, insights can be obtained regarding dependencies with social and economic phenomena. An understanding of such relationships can help governments provide better polices and implement wholesome decisions that contribute to an improved global wellness. It could help observe the ways in which various freedoms interact with one another. Data Analysis over the given measures can provide insights to help better understand patterns, interesting relations among variables, correlate attributes, and find dependencies. The project handles Data Preprocessing, Exploratory Data Analysis, and Target Prediction for the Human Freedom Score using Linear Regression. Various factors that contribute to the outcomes are analyzed from relations between the indicators that can help provide improved solutions in the future.**

Keywords—Human Freedom Index, Exploratory Data Analysis, Linear Regression, Machine Learning, Human Freedom Score (hf_score), Economic Freedom Score (ef_score), Personal Freedom Score (pf_score).

I. Introduction

Any industry requires forecasting of future trends, hypothesis of present data trends, and forming relations between indicators to improve policies, provide solution, and even improve decision making skills to come up with social, civic, and economic solutions based on them. By using data science concepts, historical and present situations and scenarios can be estimated and analyzed to make predictions, improve decisions, and help elevate world policies for the future generation.

The Human Freedom Index is an annual assessment that assesses the level of human freedom in 165 countries and territories worldwide, accounting for 98.1 percent of the global population as of 2021. It is a broad-based metric that considers both economic and personal freedom based on various attributes and specifications. The Cato Institute, the Fraser Institute, and the Heritage Foundation together publish the Human Freedom Index [1].

Published in 2021, it is by far the most extensive measurement of freedom that has been devised so far. This dataset can be analyzed and made subject to Data Analysis and Data Modelling techniques to obtain insights with respect to freedom measures, their patterns, and reasons for their trajectories. Exploratory Data Analysis (EDA) over the given Human Freedom Index measures can provide insights to help better understand patterns of the data, interesting relations among variables, correlate attributes, and find dependencies between them. Based on accuracies of various Machine Learning (ML) models, the best technique can be chosen to represent the Human Freedom Score (hf_score) based on other attributes as a target prediction. The project handles Data Preprocessing, Exploratory Data Analysis, and Target Prediction for the Human Freedom Score using Linear Regression. Numerous factors that contribute to the outcomes of EDA are analyzed from relations between the indicators that can help provide improved solutions in the future. The main objectives are four-fold:

- To understand the Human Freedom Index by analyzing statistics and describing data,
- To perform Data Preprocessing by handling missing data, obtaining a statistical overview of attributes and the target values,
- To perform Exploratory Data Analysis and observe patterns and statistical trends of attributes, analyze them, and discover correlations among attributes, and
- To predict the Personal Freedom Score, Economic Freedom Score, and finally the Human Freedom Score based on all the attributes using Linear Regression.

It also includes a study of a few ML models with the aim of selecting the model that provides the best accuracies for the Target Value analysis. Predictive Modelling is followed by a check on attribute correlations. Factors behind their past trajectories are included in the future scope of this project.

II. Related Work

Due to the tremendous volume of digital data being generated and at such a rapid pace, information cannot be easily interpreted by an individual but instead must be relied on machines to interpret and process it. Data Science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. It uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources and are presented in various formats. Extracting knowledge from data sets (Big Data), and applying the knowledge/insights to solve problems is the ultimate goal. It incorporates skills from computer science, statistics, information science, mathematics, information visualization, data integration, graphic design, and business.

These have multiple categories of processes involved like data mining, statistical analysis, data engineering, business intelligence, machine learning, deep learning, data architecture, data visualization, analysis, etc.

A. Involved Techniques

Data Mining is the process of analyzing a large batch of information to discern trends and patterns [2]. The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table. Warehousing is an important aspect of data mining. Warehousing is when companies centralize their data into one database or program. With a data warehouse, an organization may spin off segments of the data for specific users to analyze and use. However, in other cases, analysts may start with the data they want and create a data warehouse based on those specifications Data Mining techniques include Association Rules, Classification, Clustering, Decision Trees, K-Nearest Neighbor, Neural Networks, Predictive Analysis, etc. Applications are vast in areas as long as there exists data for analysis.

Statistical Analysis is the process of collecting and analyzing data in order to discern patterns and trends [3]. It is a method for removing bias from evaluating data by employing numerical analysis. This technique is useful for collecting the interpretations of research, developing statistical models, and planning surveys and studies. There are six types of statistical analysis - descriptive analysis, inferential analysis, predictive analysis, prescriptive analysis, exploratory data analysis, and casual analysis. There are in general five steps to conduct a statistical analysis. First, the data to be analyzed is identified and described. The next step is to establish a relation between the data analyzed and the sample population to which the data belongs. The third step is to create a model that clearly presents and summarizes the relationship between the population and the data. Then, the data model is proved to be either valid or invalid. Predictive Analysis is performed to predict future trends and events likely to happen.

Data Engineering refers to the building of systems to enable the collection and usage of data [4]. This data is usually used to enable subsequent analysis and data science; which often involves machine learning. Making the data usable usually involves substantial compute and storage, as well as data processing and cleaning. The design of data systems involves several components such as architecting data platforms and designing data stores. Data Modelling produces a data model, an abstract model to describe the data and relationships between different parts of the data.

Business Intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company’s activities [5]. It encompasses data mining, process analysis, performance benchmarking, and descriptive analytics. BI parses all the data generated by a business and presents easy-to-digest reports, performance measures, and trends that inform management decisions. This can reduce the need to capture and reformat everything for analysis, saving analytical time and increasing the reporting speed. Some BI tools and software include Spreadsheets, Reporting software, Data Visualization software, Data Mining tools, Online Analytical Processing (OLAP) and Online Transactional Processing (OLTP).

Machine Learning involves showing a large volume of data to a machine so that it can learn and make predictions, find patterns, or classify data. The three main Machine Learning types are Supervised, Unsupervised, and Reinforcement learning. It can also be Semi-supervised. Deep Learning (DL) is also a subset of Machine Learning.

Example of Supervised Learning algorithms include: Linear Regression, Logistic Regression, Nearest Neighbor, Gaussian Naive Bayes, Decision Trees, Support Vector Machine (SVM), Random Forest, and Unsupervised. Machine learning analyzes and clusters unlabeled datasets using machine learning algorithms. These algorithms find hidden patterns and data without any human intervention, i.e., outputs are not provided to the model. The training model has only input parameter values and discovers the groups or patterns on its own.

Some algorithms of Unsupervised Learning include K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), and Hierarchical Clustering. It is used especially when data is partially labeled and the remaining larger portion of it is unlabeled. Unsupervised techniques predict labels and feed them to supervised techniques. This technique is mostly applicable in the case of image data sets where usually all images are not labeled.

Reinforcement learning differs from supervised learning in not needing labelled input/output pairs to be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead, the focus is on finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge). Some algorithms include Temporal Difference (TD), Q-Learning, and Deep Adversarial Networks

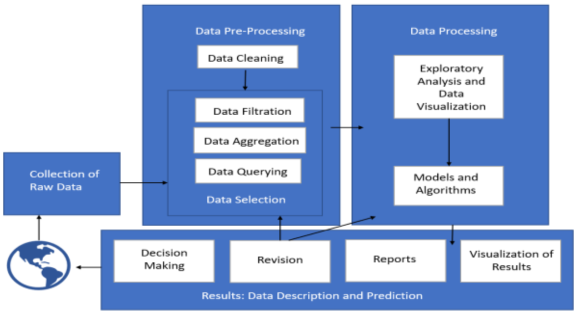
Deep learning or Deep Structured Learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised [6]. Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. Most modern deep learning models are based on artificial neural networks, specifically Convolutional Neural Networks (CNNs), although they can also include propositional formulas or latent variables organized layer-wise in deep generative models. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, convolutional neural networks and Transformers have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

The Data Architecture describes how data is managed from collection through to transformation, distribution, and consumption. It sets the blueprint for data and the way it flows through data storage systems. It is foundational to data processing operations and ML applications. The architecture design is set of standards which are composed of certain policies, rules, models and standards which manages, what type of data is collected, from where it is collected, the arrangement of collected data, storing that data, utilizing and securing the data into the systems and data warehouses for further analysis, processing the data, modelling the data, applying transformations, etc. It is dependent on Technology Requirements, Business Processing needs, Business Policies, Data Processing needs, and Objectives of the analysis. Here, after processing and visualization, Linear Regression Modelling will be applied. It is an ML Algorithm based on supervised learning that performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. This section discusses the organizational flow of phases of a Data Analysis experiment along with the underlying ML architecture.

B. *Phases of Analysis*

Analysis refers to dividing a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and subsequently converting it into information useful for decision-making. Data is collected and analyzed to answer questions, test hypotheses, or disprove theories. There are several phases that can be distinguished, described as per Figure 1.

Fig. 1. Data Analysis Process Architecture



Data Requirements - The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis (or customers, who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., individualism and cultural expression) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

Data Collection - Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data; such as Information Technology personnel within an organization. The data may also be collected from sensors in the environment, including traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

Data Processing - The phases of this intelligence cycle convert raw information into actionable intelligence or knowledge and are conceptually like the phases in data analysis. Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis, often using spreadsheet or statistical software. It can involve obtaining statistical information about attributes of the dataset.

Data Cleaning - Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that the data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers that are believed to be reliable. Unusual amounts, above or below predetermined thresholds, may also be reviewed. There are several types of data cleaning, that are dependent upon the type of data in the set; this could be phone numbers, email addresses, employers, or other values.

Quantitative data methods for outlier detection can be used to get rid of data that appears to have a higher likelihood of being input incorrectly. Textual data spell checkers can be used to lessen the amount of mis-typed words.

Exploratory Data Analysis - Once the datasets are cleaned, they can then be analyzed. Analysts apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data. The process of data exploration may result in additional data cleaning or additional requests for data; thus, the initialization of the iterative phases of the intelligence cycle. Descriptive statistics, such as, the average or median, can be generated to aid in understanding the data. Data visualization is also a technique used, in which the analyst is able to examine the data in a graphical format in order to obtain additional insights regarding the messages within the data.

Modelling and Algorithms - Mathematical formulas or models (known as algorithms) may be applied to data in order to identify relationships among the variables; for example, using correlation or causation. In general terms, models may be developed to evaluate a specific variable based on other variable(s) contained within the dataset, with some residual error depending on the implemented model's accuracy. Inferential statistics includes utilizing techniques that measure the relationships between particular variables.

Data Visualization - Visualization helps understand the results after data is analyzed. Once data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative. When determining how to communicate the results, the analyst may consider implementing a variety of data visualization techniques to help communicate the message more clearly and efficiently to the audience. Data visualization uses information displays (graphics such as, tables and charts) to help communicate key messages contained in the data. Tables are a valuable tool by enabling the ability of a user to query and focus on specific numbers; while charts (e.g., bar charts or line charts), may help explain the quantitative messages contained in the data. Based on the visualized data, relationships among attributes and correlations can be obtained.

III. Literature Survey

The main survey would be over the Human Freedom Index itself, along with any previous analysis made over the index. The two base reports include:

1. The Human Freedom Index 2021 (HFI) – A Global Measurement of Personal, Civil, and Economic Freedom, By Ian Vásquez, Fred McMahon, Ryan Murphy, and Guillermina Sutter Schneider, Published by the CATO Institute and the Fraser Institute, and
2. The Human Development Index (HDI), By Max Roser, Published by the United Nations Development Program.

The Human Freedom Index is a measure of personal, civil, and economic freedom, and it depends on several factors like (a) Personal freedom, which includes freedom of speech, religion, association, and movement, as well as security of the person and protection against physical harm, such as torture and cruel, inhuman, or degrading treatment, (b) Civil freedom, which refers to freedom of assembly and association, freedom of the press and media, and the right to participate in political processes and have a say in the decisions that affect one's life, and (c) Economic freedom, which encompasses freedom to engage in economic activities, including starting a business, trading, and investing, as well as the ability to own property and protect one's property rights.

A. Human Freedom Index

The HFI is based on data collected by several international organizations and research institutions, including the World Bank, the International Monetary Fund, and the World Health Organization. The index considers a wide range of factors, such as government regulations, taxes, trade policies, and the rule of law, to determine a country's overall level of freedom. The index is also subject to limitations and criticisms, such as measurement errors and questions about the appropriate weighting of different components of freedom.

The primary three factors which influence the Human Freedom Ranks include –

1. Historical Trends: Over the past few centuries, there has been a general trend towards greater freedom and human rights in many parts of the world. This has included the abolition of slavery, the expansion of suffrage and political rights, and the growth of free markets and private property rights,
2. Regional Differences: The level of freedom and human rights has varied significantly across different regions and countries. Some regions, such as Western Europe and North America, have generally been leaders in promoting freedom and human rights, while others, such as the Middle East and parts of Africa and Asia, have lagged behind, and
3. Recent Developments: In recent years, there has been some concern about a decline in freedom and human rights in certain countries. This has been attributed to a variety of factors, including political polarization, government repression, and the rise of authoritarian regimes.

The average human freedom rating for 165 jurisdictions in 2019 was 7.12 on a scale of 0 to 10, where 10 represents more freedom. Among the 162 jurisdictions for which we have data for 2018 and 2019, the overall level of freedom (weighing all jurisdictions equally) remained unchanged, with 82 jurisdictions decreasing their ratings and 67 countries improving. Further analysis showed that there is an unequal distribution of freedom in the world, with only 14.6

percent of the world's population living in the top quartile of jurisdictions in the HFI and 40.3 percent living in the bottom quartile. The gap in human freedom between the most-free and the least-free jurisdictions has widened since 2008, increasing by 6.6 percent when comparing the top and bottom quartile of nations in the HFI. The countries that took the top 10 places, in order, were Switzerland, New Zealand, Denmark, Estonia, Ireland, Canada and Finland (tied at 6), Australia, Sweden, and Luxembourg.

Out of 10 regions, the regions with the highest levels of freedom are North America (Canada and the United States), Western Europe, and Oceania. The lowest levels are in the Middle East and North Africa, sub-Saharan Africa, and South Asia. Women-specific freedoms, as measured by five indicators in the index, are strongest in North America, Western Europe, and East Asia and are least protected in the Middle East and North Africa, South Asia, and sub-Saharan Africa. According to the report, the top five jurisdictions in the Human Freedom Index for 2019 are Switzerland, New Zealand, Denmark, Estonia, and Ireland. The bottom five countries are, in descending order, Egypt, Sudan, Yemen, Venezuela, and Syria.

The correlation between the personal and economic freedom ratings on analysis was 0.67 for 2019. Some countries ranked consistently high in the human freedom subindexes, including Switzerland and Ireland, which ranked in the top 10 in both personal and economic freedom. By contrast, some countries that ranked high on personal freedom ranked significantly lower in economic freedom.

Of the 12 major categories that make up the index, half saw some deterioration since 2008 (comparing the 141 jurisdictions for which we have data for both 2008 and 2019). The categories of expression and information, religion, and association and assembly saw the largest decreases in freedom since 2008, while sound money saw the largest improvement.

Democracy and development are often seen as interrelated concepts, with democracy considered to be a key factor in promoting economic and social development. Democracy is often associated with economic growth and stability, as democratic systems promote transparency, accountability, and the rule of law, all of which are considered to be important for creating a favorable environment for economic growth. In democratic societies, citizens have a greater say in how resources are allocated, and markets tend to be more competitive, leading to greater efficiency and productivity. Democracy also plays a role in promoting social development by fostering greater equality, justice, and human rights. HFI is subject to several limitations, including subjectivity in measurement limitations of the data, difficulty in defining freedom, and time lag in data collection. Despite these limitations, the HFI is widely used and respected as a measure of freedom, and it provides valuable insights into the state of freedom in countries around the world.

B. Human Development Index

The HDI [7] is a composite statistic of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers of human development. It was developed by the United Nations Development Program (UNDP) as a tool to measure and compare levels of human development between countries, and to provide insights into the drivers of human development. HDI values range from 0 to 1, with higher values indicating higher levels of human development. The three key dimensions are: A long and healthy life (measured by life expectancy), access to education (measured by expected years of schooling of children at school at entry age and mean years of schooling of the adult population), and a decent standard of living (measured by Gross National Income per capita adjusted for the price level of the country).

The HDI changes over time due to changes in the underlying components, such as life expectancy, years of schooling, and gross national income per capita. These changes can be driven by a variety of factors, such as economic growth, technological advancements, and improvements in health and education systems. Life expectancy has increased significantly over the past.

The Historical Index of Human Development (HIHD) is an extension of the HDI that looks at human development over time, rather than just at a single point in time. It provides a measure of how the level of human development in a country has changed over the years and allows for a more nuanced understanding of the evolution of human development in different countries. The HIHD is constructed by estimating the HDI for each year for a given time, and then averaging those estimates to produce an average HDI value for the entire period. This allows for the comparison of human development across time, and for the examination of long-term trends in human development.

C. Limitations

Though HFI is a complex measure that seeks to capture the state of personal, civil, and economic freedom in countries around the world. However, it is subject to several limitations, including:

- The HFI is based on data collected by various international organizations and research institutions, but the interpretation of this data is subject to interpretation. The choice of parameters and their weighting can be influenced by political, cultural, and ideological biases, which can lead to differences in the way freedom is measured in different countries.
- The HFI is based on data that is not always directly comparable or of high quality across countries. This can lead to measurement errors and can result in the HFI overestimating or underestimating the level of freedom in a given country.
- Freedom is a complex and multifaceted concept that can be difficult to define and measure. The HFI seeks to capture different aspects of freedom, such as personal freedom, civil freedom, and economic

freedom, but these categories are not always clear-cut and can overlap in practice.

- The HFI is based on data that is often several years old, which can result in the index not fully capturing recent developments in freedom in a given country.
- The HFI does not take into account important contextual factors that can influence freedom, such as historical, cultural, and social factors.

IV. Supervised Learning

Supervised Machine Learning is a type of machine learning where the algorithm is trained on labeled data to make predictions or classify new, unseen data. In supervised learning, the algorithm is provided with a set of labeled examples, where the label is the target variable that the algorithm is trying to predict. As shown in Figure 2, The algorithm uses these labeled examples to learn the relationship between the input features and the target variable, and then uses this learned relationship to make predictions on new, unseen data. It is used in a variety of applications, including image classification, sentiment analysis, and predictive modeling.

The algorithms used in supervised learning include linear regression, logistic regression, decision trees, and artificial neural networks, among others. The choice of algorithm depends on the type of data, the complexity of the relationship between the features and target variable, and the desired outcome of the analysis. The most commonly used supervised machine learning algorithms include linear regression (used to model the linear relationship between a dependent variable and one or more independent variables), logistic regression (used for binary classification problems where the target variable can take on only two values), decision trees (used for classification and regression tasks, where the model makes a prediction by recursively splitting the data into smaller subsets based on the feature values), random forest (an ensemble learning method that uses multiple decision trees to make a prediction, where the final prediction is based on the average or majority vote of the individual trees, support vector machines (a linear model used for binary classification problems that finds the best hyperplane to separate the two classes of data), naive bayes (a probabilistic model used for classification tasks, where the prediction is based on the maximum likelihood estimate of the class based on the feature values), k-nearest neighbors (an instance-based learning algorithm that classifies new data based on the majority vote of its k-nearest neighbors in the training data), neural networks (a machine learning model that uses artificial neurons to learn complex relationships between the input features and target variable), and gradient boosting (an ensemble learning method that uses multiple weak models, such as decision trees, to make a prediction, where the final prediction is based on the weighted sum of the individual model predictions).

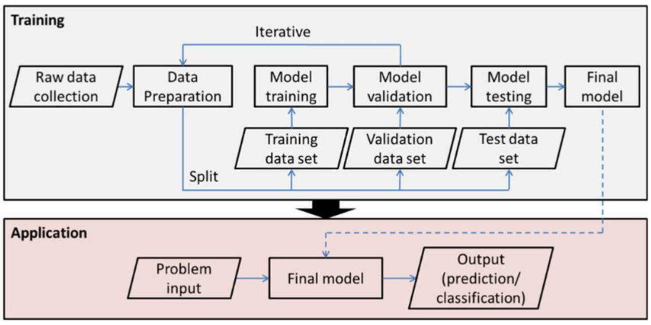


Fig. 2. Supervised Machine Learning Implementation

The aim of supervised machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. As adaptive algorithms identify patterns in data, a computer learns from the observations. When exposed to more observations, the computer improves its predictive performance. The aim of supervised machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. As adaptive algorithms identify patterns in data, a computer learns from the observations. When exposed to more observations, the computer improves its predictive performance. Supervised Algorithms are great for labelled data sets and can prove to have higher accuracies than other algorithms. A few advantages include prediction accuracy, ease of implementation, interpretability, automated feature selection, etc. Some of the disadvantages include limited to known labels, overfitting, assumes linearity, computational burden.

A. Linear Regression

Linear Regression is a statistical method used to model the linear relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as predictors or features) [8]. The goal of linear regression is to find the line of best fit that minimizes the difference between the predicted and actual values of the dependent variable. It is recommended when there is a continuous target variable and one or more continuous or categorical predictors. A clear understanding of the underlying relationships between the variables must exist such that they can be well-approximated by a linear function. A few advantages of this technique include its simplicity to implement, fast and efficient, easy to interpret, and good performance with linear relationships. Some of the drawbacks include its limited representation of non-linear relationships, assumes linearity and independence, sensitivity to outliers, and its assumption of normally distributed errors. In general, linear regression is a good starting point for simple problems with linear relationships between the features and the target variable. However, for more complex problems with non-linear relationships or for data sets with outliers, other algorithms, such as decision trees, random forests, and neural networks, may be more appropriate.

B. Logistic Regression

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification problems, where the outcome or target variable can take only two values. Logistic Regression models the relationship between the independent variables and the odds of the target variable being a certain value, using a logistic function to model the probability of the target variable. The logistic regression algorithm then learns the coefficients of the independent variables that best fit the data, allowing it to make predictions on new data. Its assumptions include the dependent variable must be categorical in nature and the independent variable should not have multi-collinearity. Its limitations include limited representation of complex relationships, Assumes independence between predictors, can be sensitive to outliers, and assumes normally distributed errors. Logistic Regression is recommended when there exists a binary target variable and one or more continuous or categorical predictors. It models the relationship between the predictors and the probability of the target variable taking on one of the two possible values.

C. Decision Trees

A decision tree is a tree-based machine learning algorithm that is used for both classification and regression tasks [9]. It is a graphical representation of the decisions and their possible consequences, including chance events, represented in the form of a tree structure. They are a popular and widely used machine learning algorithm for both classification and regression tasks. It is a tree-based model where internal nodes represent features of the data and leaf nodes represent the decisions or predictions. The main advantage of decision tree models is that they are simple to understand and interpret, and they can handle both categorical and numerical data. The main disadvantage is that they can easily overfit the training data, leading to poor generalization performance on new data. To mitigate this, techniques such as pruning, bagging, and random forests can be used to improve the decision tree model's performance. Interoperability, its ability to handle both continuous and categorical data are its advantages. Its drawbacks include overfitting, instability, unbalanced classes, greedy algorithm, and bias towards features with many outcomes.

Decision Trees have several disadvantages that can limit their performance, leading to the development of alternative models such as Random Forests as mentioned in the previous section. To address these limitations, the Random Forest algorithm was developed. Random Forests are an ensemble learning method that uses multiple decision trees and combines their predictions to make a final prediction. By aggregating the predictions of multiple trees, Random Forests can reduce the variance and instability of individual trees and provide more accurate predictions. Additionally, the use of random subsets of features and samples during tree construction can reduce overfitting and improve the generalization performance of the model. Advantages of Random Forest Modeling include improved accuracy, reduced variance, handling complex relationships, handling non-linear relationships, handling missing data, and interpreting results. Some of its disadvantages include its computational costs, uninterpretable model, overfitting, and bias.

D. Support Vector Machine

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm used for classification and regression analysis. The main idea behind SVMs is to find a boundary that separates the data points into different classes, known as the "maximum margin hyperplane." The maximum margin hyperplane is the one that has the largest distance between the boundary and the closest data points, known as "support vectors." These are used in a wide range of applications, including text classification, image classification, and bioinformatics. They are particularly well-suited for problems where there are many features or the data is not linearly separable. Its advantages include improved accuracy, handling non-linear relationships. Handling high-dimensional data, handling imbalanced data, handling noise data, interpreting results, and versatility.

E. Naïve Bayes

Naive Bayes is a machine learning algorithm based on Bayes' theorem, which states that the probability of an event is equal to the prior probability of the event multiplied by the likelihood of the event given the evidence [10]. The "naive" part of the name comes from the assumption that the features in the data are independent of each other, which is often not the case. Advantages include simplicity, fast and efficient, good performance with small data sets, handles irrelevant features well. Disadvantages include naïve assumptions, limited representation of complex relationships, reliance on probability estimates, and limited to binary or multiclass problems.

F. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a simple, non-parametric machine learning algorithm that can be used for both classification and regression problems [11]. It is based on the idea that the target variable for a given data point can be predicted based on the target variable of its nearest neighbors by Euclidian distance. Its simplicity to understand and implement, versatility, and no data assumptions are its advantages. However, its high computational costs, sensitivity to outliers, inability to work well with categorical variables, and poor performance in high-dimensional space are its drawbacks.

V. Proposed Solution

This project includes Exploratory Data Analysis over the Human Freedom Index (HFI) and Target Prediction of the Human Freedom Score (hf_score), Personal Freedom Score (pf_score), and Economic Freedom Score (ef_score) based on the data set for the time period from 2008 to 2019 for 165 countries. The project uses Supervised Algorithms since the records are all labelled with their scores. Of these algorithms, Linear Regression Modelling is performed over the target values.

Exploratory Data Analysis (EDA) over the given Human Freedom Index measures can provide insights to help better understand patterns of the data, interesting relations among variables, correlate attributes, and find dependencies between them. Python code handles Data Preprocessing, Exploratory Data Analysis, and Target Prediction for the Human Freedom Score using Linear Regression. Various factors that contribute to the outcomes of EDA are analyzed from relations between the indicators that can help provide improved solutions in the future.

A. Methodology

The different processes performed on the dataset is depicted in the above timeline Figure 3. It broadly consists of the below phases.

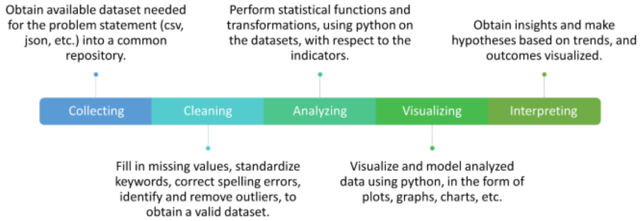

Obtain available dataset needed for the problem statement (csv, json, etc.) into a common repository.
Perform statistical functions and transformations, using python on the datasets, with respect to the indicators.
Obtain insights and make hypotheses based on trends, and outcomes visualized.
Collecting
Cleaning
Analyzing
Visualizing
Interpreting
Fill in missing values, standardize keywords, correct spelling errors, identify and remove outliers, to obtain a valid dataset.
Visualize and model analyzed data using python, in the form of plots, graphs, charts, etc.

Fig. 3. Implementation Timeline

1. Data processing by handling missing data by filling NaN values., obtaining an overview of the data by describing parameters and obtaining their Value Counts, and observing target value statistics by describing the target attributes.
2. Exploratory Data Analysis by observing Target trends by plotting their line plots, analyzing the attribute patterns by plotting each attribute over the timeframe for each country and/or region, and calculating positive and negative correlations between attributes by plotting a Heat Map.
3. Target Prediction of Personal, Economic, and Human Freedom Scores by building a Linear Regression Model given the training data, by predicting the Test Data over the Model, calculating the Model Accuracy, plotting the Residual Scatter Graph for the model, and obtaining the Intercept and Coefficients of the model to build an equation for the scores.

B. Data Set

A brief list of the attributes involved and analyzed in the dataset include (Personal and Economic Freedom) the below tables. The collected data exists in a Comma Separated Values (CSV) format.

1. Rule of Law	pf_rol
i. Procedural Justice	pf_rol_procedural
ii. Civil Justice	pf_rol_justice
iii. Criminal Justice	pf_rol_criminal
2. Security and Safety	pf_ss
i. Homicide	pf_ss_homicide
ii. Disappearances, Conflicts, and Terrorism	pf_ss_disappearances
a. Disappearances	pf_ss_disappearances_disap
b. Violent Conflicts	pf_ss_disappearances_violent
c. Organized Conflicts	pf_ss_disappearances_organized
d. Terrorism Fatalities	pf_ss_disappearances_fatalities
e. Terrorism Injuries	pf_ss_disappearances_injuries
f. Freedom from Torture	pf_ss_disappearances_torture
g. Freedom from Political killings	pf_ss_killings
3. Movement	pf_movement
i. Freedom of Movement (V-Dem)	pf_movement_vdem
a. Freedom of foreign movement	pf_movement_vdem_foreign
b. Freedom of movement for men	pf_movement_vdem_men
c. Freedom of movement for women	pf_movement_vdem_women
ii. Freedom of Movement (CLD)	pf_movement_cld
4. Freedom of Religion	pf_religion
i. Freedom of Religion	pf_religion_freedom
a. Freedom of religion (V-Dem)	pf_religion_freedom_vdem
b. Freedom of religion (CLD)	pf_religion_freedom_cld
ii. Religious Organisation and Repression	pf_religion_suppression
5. Association, Assembly, and Civil Society	pf_assembly
i. Civil Society Entry and Exit	pf_assembly_entry
ii. Freedom of Assembly	pf_assembly_freedom
a. Freedom of assembly (Freedom House)	pf_assembly_freedom_house
b. Freedom of assembly (BTI)	pf_assembly_freedom_bti
c. Freedom of assembly (CLD)	pf_assembly_freedom_cld
iii. Freedom to form and run Political Parties	pf_assembly_parties
a. Barriers to parties	pf_assembly_parties_barriers
b. Party bans	pf_assembly_parties_bans
c. Opposition parties autonomy	pf_assembly_parties_auton
iv. Civil Society Repression	pf_assembly_civil
6. Expression and Information	pf_expression
i. Press Killed	pf_expression_killed
ii. Press Jailed	pf_expression_jailed
iii. Freedom of Academic and Cultural Expression	pf_expression_cultural
iv. Harassment of Journalists	pf_expression_harass
v. Government Censorship Effort	pf_expression_gov
vi. Internet Censorship Effort	pf_expression_internet

vii. Media Self-censorship	pf_expression_selfcens
viii. Media Freedom	pf_expression_media
ix. Freedom of Expression a. Freedom of expression (BTI) b. Freedom of expression (CLD)	pf_expression_freedom pf_expression_freedom_bti pf_expression_freedom_cld
7. Relationships	pf_identity
i. Same-sex Relationships a. Male-to-male relationships b. Female-to-female relationships	pf_identity_same pf_identity_same_m pf_identity_same_f
ii. Divorce	pf_identity_divorce
iii. Inheritence Rights	pf_identity_inheritance
iv. Female Genital Mutilation	pf_identity_fgm

Table 1. Personal Freedom Attributes

1. Size of Government	ef_government
i. Government Consumption	ef_government_consumption
ii. Transfers and Subsidies	ef_government_transfers
iii. Government Investment	ef_government_enterprises
iv. Top Marginal Tax Rate a. Top marginal income tax rate b. Top marginal income and payroll tax rates	ef_government_tax ef_government_tax_income ef_government_tax_payroll
v. State Ownership of Assets	ef_government_soa
2. Legal System and Property Rights	ef_legal
i. Judicial Independence	ef_legal_judicial
ii.Impartial Courts	ef_legal_courts
iii. Protection of Property Rights	ef_legal_protection
iv. Military Interference in Rule of Law and Politics	ef_legal_military
v. Integrity of Legal System	ef_legal_integrity
vi. Legal Enforcement and Contracts	ef_legal_enforcement
vii. Regulatory Costs of the Sale of Real Property	ef_legal_regulatory
viii. Reliability of Police	ef_legal_police
3. Sound Money	ef_money
i. Money Growth	ef_money_growth
ii. Standard Deviation of Inflation	ef_money_sd
iii. Inflation: Most Recent Year	ef_money_inflation
iv. Freedom to own Foreign Currency Bank Accounts	ef_money_currency
4. Freedom to Trade Internationally	ef_trade
i. Tariffs a. Revenue from trade taxes (% of trade sector) b. Mean tariff rate c. Standard deviation of tariff rates	ef_trade_tariffs ef_trade_tariffs_revenue ef_trade_tariffs_mean ef_trade_tariffs_sd
ii. Regulatory Trade Barriers a. Non-tariff trade barriers b. Compliance costs of importing and exporting	ef_trade_regulatory ef_trade_regulatory_nontariff ef_trade_regulatory_compliance
iii. Black-market Exchange Rates	ef_trade_black
iv. Controls of Movement of Capital and People a. Financial openness b. Capital controls c. Freedom of foreigners to visit	ef_trade_movement ef_trade_movement_foreign ef_trade_movement_capital ef_trade_movement_visit
5. Regulation	ef_regulation
i. Credit Market Regulations a. Ownership of banks b. Private-sector credit c. Interest rate controls/negative real interest rates	ef_regulation_credit ef_regulation_credit_ownership ef_regulation_credit_private ef_regulation_credit_interest
ii. Labor Market Regulations a. Hiring regulations and minimum wage b. Hiring and firing regulations c. Centralized collective bargaining d. Hours regulations e. Mandated cost of worker dismissal f. Conscription	ef_regulation_labor ef_regulation_labor_minwage ef_regulation_labor_firing ef_regulation_labor_bargain ef_regulation_labor_hours ef_regulation_labor_dismissal ef_regulation_labor_conscription
iii. Business Regulations a. Administrative requirements b. Bureaucracy costs c. Starting a business d. Impartial public administration e. Licensing restrictions f. Cost of tax compliance	ef_regulation_business ef_regulation_business_adm ef_regulation_business_bureaucracy ef_regulation_business_start ef_regulation_business_bribes ef_regulation_business_licensing ef_regulation_business_compliance

Table 2. Economic Freedom Attributes

C. Data Preparation

The HFI dataset is analyzed in the form of a DataFrame. Information regarding its size, data types, attributes observed, etc. are described. This is performed using the NumPy and Pandas libraries using the Python language. It is a data structure that organizes data into a 2-dimensional table of rows and columns, much like a spreadsheet. Hence the data set is not in tabular format. It can be accessed by calling the DataFrame object into which the data set is loaded.

From the results, it was seen that the count was not complete for all columns, as there exists missing values that must be handled. This is a part of the Data Cleaning process. The DataFrame may have Not A Number (NaN) values, redundant values, etc. that must be managed to avoid calculation errors and incompatibilities. They can be aggregated by computing the exact number of missing values in each column. It is observed that multiple columns have missing values. Missing values must be filled in by grouping existing attribute values and finding their means.

A Data Overview can be obtained by analyzing the type of data present. This includes Categorical Data columns, number of entries, their value counts, etc. Such information can help understand how viable or clean the available data is. Target Value statistics can also be obtained to do the same. Value counts can also be obtained for numerical attributes. The target value statistics for hf_score, pf_score, and ef_score can be obtained by performing describe( ) over the respective columns in the Data Frame. These have been tabulated in Table 3.

Measure	hf_score	pf_score	ef_score
count	1980	1980	1980
mean	7.117118	7.310538	6.83904
std	1.202637	1.554443	0.945897
min	3.49	2.45	2.67
25%	6.25	6.1075	6.23
50%	7.16	7.53	6.93
75%	8.1225	8.5825	7.56
max	9.15	9.67	9

Table 3: Target Value Statistics

D. Exploratory Data Analysis and Observations

The HFI contains multiple attributes based on category (Economic or Personal), region, and even country. Totally, 125 columns are existent. Initially, Target Trends can be analyzed by plotting the hf_score based on these categories. Individual line plots for each country can be obtained to track the growth of the hf_score for that country. These trajectories would help form hypothesis as to political crisis, government rule and policies, etc. for that country. This can be performed by plotting a Line Graph for the attribute using the Seaborn lineplot( ) function. The data passed is the DataFrame, the x-axis represents the year, while the y-axis represents the hf_score. The countries are identified by their respective hue colours.

The following enlist a sample of the results of region-wise comparison on the Human Freedom Score (hf_score).

- Middle Eastern countries seem to have lower hf_score, around 6. They have reduced with time.
- Oceana too seems to drop. Remaining countries have remained somewhat steady.
- Sub Saharan Africa have second lowest scores around 6.3.
- Caucasus and Central Asian countries, start at a score around 6.8 and have increased to 7.0.
- Latin America and the Caribbean have scores around 7.25, and have reduced with time, by a small extent.
- Eastern European countries of scores around 7.75 have remained steady throughout the timeline.
- East Asian countries have a score around 7.8, which has increased with time.
- Oceana starts at a score of 8.0, which has decreased with time, crossing East Asia in 2011, and drops to around 7.8.
- Western European Countries, have the second highest score of just below 8.5, and have remained steady.
- North America has the highest score above 8.5, with a few fluctuations.

The same can be performed for Personal Freedom Score (pf_score) and Economic Freedom Score (ef_score). A sample of the observations of pf_score trends are as below:

- Middle Eastern and North African countries seem to have lower pf_score, around 5.5, which has declined toward 5.0.
- Sub Saharan Africa and South Asia have the next lowest scores around 6.5, and also seem to decline with time.
- Caucasus and Central Asian countries, start at a score around 6.6 and have fluctuated to reach a maximum in 2010, a minimum in 2015, and has later increased to a bit above 6.6.
- Latin America and the Caribbean have scores around 7.5, and have reduced with time to a score below 7.5.
- East Asian countries have a score around 8.2, which has remained steady with time.
- Eastern European countries of scores around 8.2 have decreased with time, to a score around 8.0.
- Oceana starts at a score of 8.4, which has decreased to reach a minimum in 2010, and has increased to reach a maximum in 2016, and has then steadily reduced.

- North America has the second highest score just below 9.0, which has remained quite steady.
- Western European Countries, have the highest score just above 9.0, and have reduced a bit towards 9.0 with time.

A sample of the observations for ef_score include:

- Sub Saharan Africa started with the lowest ef_score scores around 6.0, which has increased with time.
- Middle Eastern and North African countries seem to have the next lowest just below 6.5, which has declines toward 6.0, likely to meet Sub Saharan Africa after 2018.
- Caucasus and Central Asian countries, start at a score juts below 7.0, which has faced many fluctuations to eventually increase to a score just above 7.0.
- Eastern European countries of scores just above 7.0 have remained steady, with a slight increase.
- East Asian countries have a score just below 7.5, which has increased with time, to a score towards 7.7.
- Western European countries have the second highest score at around 7.75, which has fluctuated to a minimum in early 2009, a maximum in 2014, which has remained quite steady.
- North America has the highest score just that starts at 8.25, which has fluctuated to reach a minimum in 2011, a maximum in 2017, and has reduced a bit towards 8.0 with time.

The rankings of each country can be obtained by sorting their score values. Each attribute values can be plotted against each year region wise and country wise to find their patterns of growth or decrease. From these results, it can be observed that New Switzerland and New Zealand have the maximum hf_score of 9.14. Syrian Arab Republic has the lowest hf_score around 3.66.

Few countries have significant variations. Others not so much. To get a clearer view, the attribute can be plotted region-wise. Here, the hue is identified for each region. The rankings of each country can be obtained by sorting their score values. Attribute analysis can be performed by plotting each attribute against each year, say region-wise. Further, attribute correlation is performed for each indicator using a Heat Map. This displays Positive Correlations and Negative Correlations among them. Positive Correlations are identified when both indicators move in the same direction, i.e., are proportional. Negative Correlations are identified when the indicators move in opposite directions, i.e., are inversely proportional.

Dependencies among each individual attribute to the other can be obtained by plotting Pair plots among the attributes. To do so, the list of indicators to be checked is obtained, excluding ranks and quartiles. Hence, correlation is checked over all numerical attributes except for hf_rank, hf_quartile, pf_rank, and ef_rank. Pair Plots can be plotted for each of the attributes, against each other. A Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It allows to plot pairwise relationships between variables within a dataset, helping to understand the data by summarizing a large amount of data in a single figure.

E. Target Predication

Initially, the DataFrame is split into Test Sets and Training Sets for Personal Freedom related attributes and Economic Freedom related attributes. The best Train Size is found for the Linear Regression model after scaling each entity of the DataFrame. Scaling is the process of standardizing the independent features present in the data in a fixed range. It is performed by the StandardScaler( ) of the SciKit-Learn Preprocessing library. The best Train size is found to be 0.78 for the pf_score data set. It gives a Train accuracy of 0.999996 and a Test accuracy of 0.9998. The split Train and Test data are separated into their independent and dependent variables X and Y respectively. The independent-X data is scaled and described. Similarly, the ef_score and the hf_score data sets are split at ratios 0.80 and 0.78, respectively.

The Train Data is fit onto the model and accuracy scores are obtained for both the Train Data and Test Data. The Root Mean Square (RMS) error is calculated for the made predictions. Its ideal value is 0.00. To do so, the LinearRegression( ) from the SciKit-Learn Linear Model library is used. The Train data is fit into the model. The R-square score mentioning the goodness of fit is calculated for the Train Data and Test Data. Both show good accuracies of around 99.9%. The RMS error is calculated as mentioned, and is found to be 0.003 for the Train data set and 0.034 for the Test data set. . The fittings are depicted in Figure 3, Figure 4, and Figure 5 for each of the scores.

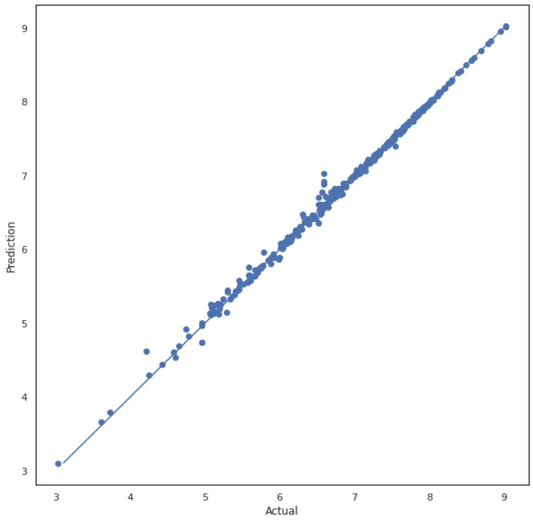


Fig. 3. Personal Freedom Score Prediction

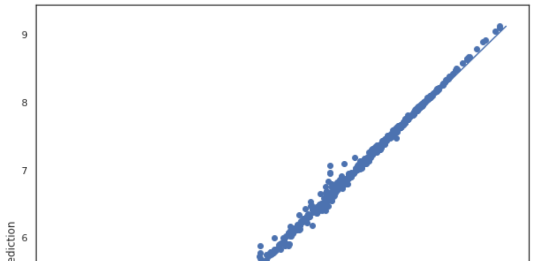


Fig. 4. Economic Freedom Score Prediction

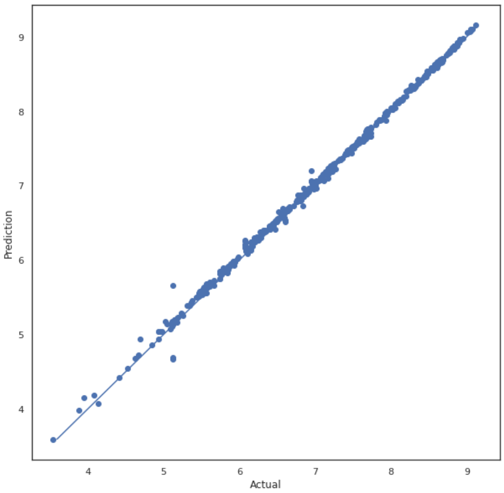


Fig. 5. Human Freedom Score Prediction

The predicted values from the Linear Regression model are plotted against the actual values. This is performed by plotting a Scatter Plot for the Test data and the Predicted data. This is done by using scatter( ) of Matplotlib library. They are found to be linear. The equation parameters, i.e., coefficients and intercepts of the liner model can be identified by accessing coef_ and intercept_ variables of the obtained Linear Model.

Additionally, residuals can be plotted as a Distribution Plot for Residual Analysis. This is necessary to validate the Regression Model obtained. If the error term in the model satisfies the assumptions of Ordinary Least Squares (OLS), then the model is considered to be valid. An ideal Residual Plot must be normalized. If it is bimodal, it indicates the dependency of a binary variable. On the other hand, this may be due to inconsistent ranges of outputs for discrete input values of the attribute. This is performed for pf_score and ef_score. For both, the Residual plot tends to be normalized as seen in Figure 6, Figure 7, and Figure 8.

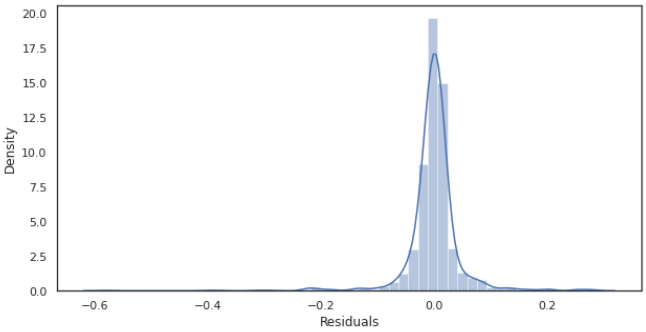


Fig. 6. Personal Freedom Score Residual Plot

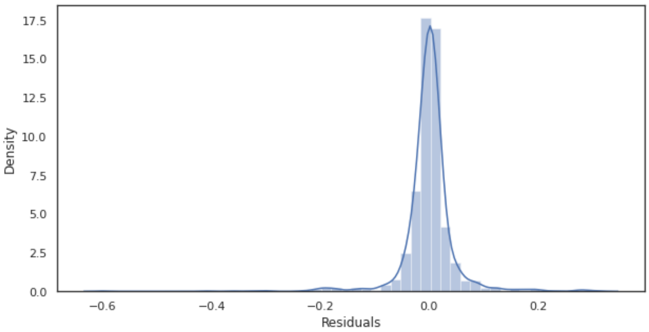


Fig. 7. Economic Freedom Score Residual Plot

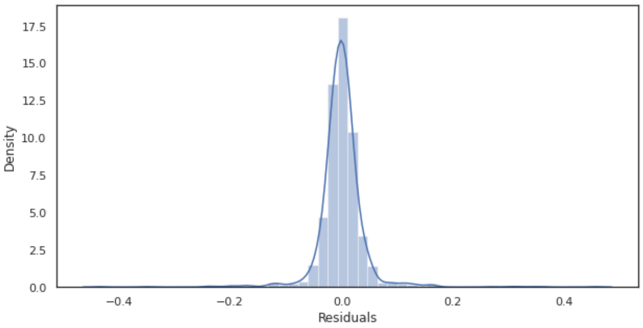


Fig. 8. Human Freedom Score Residual Plot

From the Residual plot of hf_score, it can be observed that it is dependent on pf_score and ef_score. Hence, a bimodal Scatter Plot is obtained. Hence, these attributes are to be removed from the Regression Equation. The modelling is performed again to obtain a normalized Residual plot. Hence, the scores have each been modelled to an accuracy of 99.9%.

F. Performance Metrics

The metrics listed under Table 4 are calculated based on their respective methods from SciKit-Learn library. All calculations performed use the SciKit-Learn Metrics library.

Parameters	**Train Data**	**Test Data**	**Desirable Range**
Accuracy Scores	0.9987	0.9977	>0.70
Precision	-	-	-
Recall	-	-	-
F1-Score	-	-	-
Mean Absolute Error (MAE)	0.0237	0.0439	0.00^{+}
Mean Squared Error (MSE)	0.0018	0.0043	0.00^{+}
R-Squared	0.9987	0.9969	1.00^{-}

Table 4: Performance Measures for hf_score Prediction

The confusion matrix is used to tell how many predictions were classified correctly or incorrectly. Since this project uses a Regression Model, a continuous output is obtained. Classification does not take place. Hence, Precision, Recall, and F1-Score are ruled out.

G. *Improvisations*

Cross Validation results from the remaining Supervised algorithms. The best algorithm can be chosen based on their accuracy scores and modelled further for improved findings. It is a technique in statistical modeling that is used to evaluate the performance of a model. The goal of cross-validation is to estimate how well a model will generalize to unseen data, that is, how well the model will perform on new, previously unseen data.

Insights can also be found specific to a certain country or region. This would improve motivational specific analysis into the Human Freedom Index. For example, a county like India can be analyzed on its parameters and their factors based on the Indian economy, government policies, law, rules, and regulations. This can help make predictions, improve decisions, and help elevate Indian policies for the future generation. Additionally, the observed results and insights can be justified and analyzed.

VI. Conclusion

Data analysis is a process of systematically examining and interpreting data to extract insights and knowledge that can be used to make informed decisions. It plays a critical role in solving real-world problems by providing decision makers with a clearer understanding of the problem, the data related to it, and the potential solutions.

For example, the Human Freedom Index data can be analyzed to better understand the factors that contribute to greater levels of personal and economic freedom in various countries. This analysis can help identify best practices and policies that could be adopted by governments to improve human freedom and well-being. Additionally, data analysis can help identify areas where further research is needed to address specific challenges and obstacles to human freedom. By providing a more comprehensive and detailed understanding of the state of human freedom, data analysis can support efforts to promote and protect human rights and freedom around the world. This is what has been developed and implemented.

The objective of this project included understanding the Human Freedom Index by analyzing statistics and describing data; performing Data Preprocessing by handling missing data, obtaining a statistical overview of attributes and the target values; performing Exploratory Data Analysis and observe patterns and statistical trends of attributes, analyze them, and discover correlations among attributes; and predicting the Personal Freedom Score, Economic Freedom Score, and finally the Human Freedom Score based on all the attributes using Linear Regression. These have been successfully performed. Additionally, the model was assessed for its performance measures by a variety of parameters.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this paper.

Declaration Of Competing Interests

The author(s) report there are no competing interests to declare.

Availability of Data and Materials

The datasets generated and/or analyzed during the current study are available publicly published by the CATO Institute at https://www.cato.org/human-freedom-index/2021.

Authors' Contributions

Sahana Ashok carried out analysis of literature papers on the Human Freedom Index and existing works. Data preparation, cleaning, exploratory data analysis, and target prediction by Linear Regression using Python code was implemented. Further analysis was made into observations and performance metrics was calculated. Overall analysis of the results was documented. Raj Bharath supervised and guided Sahana Ashok over the implementation phases and documentation steps. Reviews were conducted regularly.

References

[1] Ian Vásquez, Fred McMahon, Ryan Murphy, and Guillermina Sutter Schneider, The Human Freedom Index 2021: A Global Measurement of Personal, Civil, and Economic Freedom (Washington: Cato Institute and the Fraser Institute, 2021).

[2] Twin, A. (2023, January 20). What is data mining? how it works, benefits, techniques, and examples. What Is Data Mining? How It Works, Benefits, Techniques, and Examples. Retrieved January 21, 2023, from: 'https://www.investopedia.com/terms/d/datamining.asp'

[3] Thomson Reuters. 2013. Computational Statistics. 2012 Journal Citation Reports. Web of Science (Science ed.).

[4] Wikimedia Foundation. (2023, January 9). Data Engineering. Retrieved February 1, 2023, from: 'https://en.wikipedia.org/wiki/Data_engineering'

[5] Baars, H, Kemper, HG. Management Support with Structured and Unstructured Data - an Integrated Business Intelligence Framework.

[6] Information Systems Management 2008; 25:2. 132-148.

[7] Wikimedia Foundation. (2023, February 5). Deep learning. Deep Learning. Retrieved February 6, 2023, from: 'https://en.wikipedia.org/wiki/Deep_learning'

[8] Roser, M. (2014). Human Development Index (HDI). Our World In Data. Retrieved December 10, 2022, from: 'https://ourworldindata.org/human-development-index'

[9] Swaminathan, S. (2019, January 18). Linear regression. Linear Regression - Detailed View. Retrieved December 10, 2022, from: 'https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86'

[10] IBM. (2021, August 3). Decision Tree Models. IBM Documentation. Retrieved December 15, 2022, from: 'https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=trees-decision-tree-models'

[11] Ray, S. (2023, February 6). Learn Naive Bayes Algorithm: Naive Bayes Classifier Examples. Naive Bayes Explained. Retrieved February 6, 2023, from: 'https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained'

[12] Harrison, O. (2019, July 14). Machine learning basics with the K-nearest neighbors algorithm. K-Nearest Neighbors. Retrieved December 10, 2022, from: 'https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761'

[13] Alicia L. Carriquiry. (2015). Measurement Error Models. International Encyclopedia of the Social & Behavioral Sciences (Second Edition).

[14] Gerda Claeskens, Maarten Jansen. (2015). Model Selection and Model Averaging. International Encyclopedia of the Social & Behavioral Sciences (Second Edition).

[15] Kumari K, Yadav S. Linear regression analysis study. J Pract Cardiovasc Sci 2018;4:33-6.

[16] Anscombe, F. and Tukey, J. W. (1963), The Examination and Analysis of Residuals, Technometrics, pp. 141-160.

[17] Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley and Sons.

[18] Wilk, M. B. and Gnanadesikan, R. (1968), Probability Plotting Methods for the Analysis of Data, Biometrika, 5(5), pp. 1-19.

[19] Chatfield, C. (1989). The Analysis of Time Series: An Introduction, Fourth Edition, Chapman & Hall, New York, NY.

[20] Evans, Hastings, and Peacock (2000), Statistical Distributions, 3rd. Ed., John Wiley and Sons.

[21] Mosteller, Frederick and Tukey, John (1977), Data Analysis and Regression, Addison-Wesley.

[22] Snedecor, George W. and Cochran, William G. (1989), Statistical Methods, Eighth Edition, Iowa State University Press.

[23] Hill T, Lewicki P (2006) Statistics: methods and applications: a comprehensive reference forscience, industry, and data mining. StatSoft, Inc., Tulsa.

[24] Viv Bewick, Liz Cheek, and Jonathan Ball. (2003). Statistics review 7: Correlation and regression. Critical care.

[25] Dr Ossama Embarak, Embarak, and Karkal. (2018). Data analysis and visualization using python. Springer.

[26] Nils Gehlenborg and Bang Wong. (2012). Heat maps. Nature Methods.

[27] Michel Jambu. (1991). Exploratory and multivariate data analysis. Elsevier.

[28] Fabio Nelli. (2015). Python data analytics: Data analysis and science using PANDAs, Matplotlib and the Python Programming Language. Apress.

[29] Diamond, M., & Mattia, A. (n.d.). Data visualization: An exploratory study into the software

[30] tools used by businesses. Journal of Instruction Pedagogies, 17, 1–7.

[31] Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1–23.

[32] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, and Subhendu Kumar Pani. (2019). Exploratory data analysis using python. International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[33] Guido Van Rossum et al. (2007). Python programming language. In USENIX annual technical conference.