WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI-Powered Crop Yield Prediction Using Machin Learning

Mr.M. Mahesh[1], Mr.M. Jagadeesh Kumar[2], Mr.K.M.Tejewar[3], Mr.M.Buddanna[4], Mr.M.E. Palanivel[5], Dr.R.Karunia Krishnapriya[6], Mr. E.Purushotham[7]

[1,2,3,4] B.Tech student, [5]Professor, [6]Professor, [7]Professor
Sreenivasa Institute of Technology and Management Studies, Chittoor, India.

Abstract

This research suggests a novel methodology for the prediction of crop yield utilizing the strength of artificial intelligence (AI) and machine learning (ML) methods. By leveraging advanced algorithms and large datasets, our model seeks to precisely predict crop yields, supporting farmers, policymakers, and stakeholders in making informed choices. By combining AI technologies, we tackle the intricacies involved in agricultural systems, such as environmental heterogeneity, soil type, and pest pressure. Our approach employs ML algorithms to examine past yield data, weather patterns, soil type, and other pertinent factors to produce accurate predictions for future crop yields. The suggested AI-based crop yield prediction framework presents a potential means of improving agricultural productivity, maximizing resource utilization, and reducing crop production-related risks.

Keywords: Crop yield prediction, artificial intelligence, machine learning, agriculture, precision agriculture

Introduction

Farming is a crucial part of the sustenance and economic growth of numerous countries, so optimizing crop yields becomes an important task. Conventional methods of farming, although effective to some extent, depend mostly on the skill of farmers and are prone to uncontrolled environmental factors. Now that advanced technologies exist, the potential for revolutionizing agriculture using data-based solutions is increasing. Of these, machine learning is a potent tool for making more accurate and reliable predictions of crop yields.

Machine learning, a branch of artificial intelligence, focuses on the application of statistical models and algorithms to analyze and interpret large data sets. In agriculture, it allows for the processing of huge volumes of data concerning weather patterns, soil conditions, crop cultivation practices, and past yield trends. Drawing on these sources of data, machine learning algorithms can detect patterns and correlations that may not be visible to human analysts. Implementation of machine learning to predict crop yield has numerous benefits. It can integrate real-time information, which can be used for dynamic realignment of farm activities. Predictive models also can offer hints regarding the optimum utilization of water, fertilizers, and pesticides, thus boosting productivity as well as sustainability. Besides, the accurate prediction of yield can support farmers in informed decisions regarding planning for crops, marketing, and risk management.

While its potential is promising, machine learning for crop yield prediction is not without its challenges. Prediction accuracy varies with the availability and quality of data, which can prove heterogeneous across regions and crops. The intricate nature of agricultural ecosystems, subject to an unimaginable number of biotic and abiotic factors, calls for complex modeling techniques and iterative algorithm refinement.

On that note, in this scenario, this project sets out to construct a robust crop yield prediction model using machine learning, emphasizing on combining heterogeneously distributed data sources and sophisticated algorithmic solutions. In line with confronting challenges and realizing opportunities of machine learning, the project hopes to work towards enhancing farm practices by supporting food security as well as sustainable development.

INTRODUCTION

Data Science

Data science is a multidisciplinary field that employs scientific processes, algorithms and systems, processes and scientific methods to derive insights and knowledge from structured and unstructured data, and deploy knowledge and actionable insights derived from data to a wide variety of application domains.

The word "data science" has been traced to 1974, when Peter Naur suggested it as a synonym for computer science. In 1996, the International Federation of Classification Societies was the first conference to feature data science as a topic specifically. But the definition was still not fixed.

The term “data science” was first coined in 2008 by D.J. Patil, and Jeff Hammer bacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.

Data science is the scientific study that mingles domain know-how, coding proficiency, and mathematics and statistical knowledge to develop meaningful insights out of data.

Data science can be termed as a combination of business savviness, mathematics, tools, machine learning methods and algorithms, all of which assist us in discovering the underlying patterns or insights from unstructured data which can be of huge use in the creation of huge business decisions.

Data Scientist:

Data scientists analyze what questions should be answered and where the corresponding data can be found. They possess business skills and analytical skills and also the capability to mine, clean, and display data. Companies employ data scientists to procure, handle, and analyze large volumes of unstructured data.

Skills Needed for a Data Scientist:

Programming: Python, SQL, Scala, Java, R, MATLAB.

Machine Learning: Natural Language Processing, Classification, Clustering.

Data Visualization: Tableau, SAS, D3.js, Python, Java, R packages.

Big data platforms: MongoDB, Oracle, Microsoft Azure, Cloudera.

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is the imitation of human intelligence in machines that learn to think like humans and replicate their behavior. The term can also be used to refer to any machine that shows characteristics of a human mind, such as learning and solving problems.

Artificial intelligence (AI) refers to intelligence exhibited by machines, as compared to the natural intelligence shown by humans or animals. Top AI textbooks characterize the discipline as the study of "intelligent agents" any system that senses its environment and acts in a way that maximizes its probability of success at its goals. Some common accounts employ the term "artificial intelligence" to refer to machines that simulate "cognitive" functions that humans attribute to the human mind, like "learning" and "problem solving", but this definition is not accepted by leading AI researchers.

Artificial intelligence refers to the imitation of human intelligence processes by machines, particularly computer systems. Some examples of AI applications are expert systems, natural language processing, and speech recognition and machine vision.

Applications of AI include sophisticated web search engines, recommendation systems (employed by YouTube, Amazon and Netflix), Comprehension of human speech (like Siri or Alexa), autonomous vehicles (like Tesla), and playing at the world's top level in strategic game systems (like chess and Go), As computers get more powerful, tasks deemed to need "intelligence" are usually taken out of the definition of AI, a phenomenon referred to as the AI effect. For example, optical character recognition has often been left out of what would be thought of as AI, having fallen into common use.

Artificial intelligence was established as an academic field in 1956 and has since had a number of waves of optimism, followed by disappointment and the withdrawal of funds (an "AI winter"), followed by fresh approaches, success and renewed investment. Research on AI has experimented and abandoned a large number of disparate methods in its existence, ranging from simulation of the brain to modeling of human problem solving, formal logic, large knowledge bases and emulating animal behaviour. In the early decades of the 21st century, extremely mathematical statistical machine learning has prevailed field, and this methodology has been enormously successful, contributing to the resolution of numerous problem cases across industry and academia.

The multiple sub-fields in AI research all focus on specific aims and applications of specific tools. The early classic aims for AI research have included reasoning, knowledge representation, planning, learning, natural language processing, perception and the power of movement and manipulation of objects. General intelligence (the capacity to solve an arbitrary problem) is one of the long-term goals of the field. To solve such problems, AI researchers apply variants of search and mathematical optimization, formal logic, artificial neural networks, and statistical, probabilistic, and economic methods based on statistics, probability and economics. AI also borrows from computer science, psychology, linguistics, philosophy, and many other disciplines.

The discipline was established on the belief that human intelligence "can be so accurately described that a machine can be constructed to simulate it". This raises philosophical concerns regarding the mind and ethics of making artificial beings with human-like intelligence. Such questions have been probed by myth, fiction and philosophy throughout history. Science fiction and futurology have also proposed that, with such great potential and capability, AI can become a threat to the existence of humans.

As the bandwagon on AI has gained pace, vendors have been rushing to tout how their offerings employ AI. Too often what they claim as AI is merely one building block of AI, e.g., machine learning. AI depends on a body of customized hardware and software to write and train machine learning algorithms. There isn't a single programming language associated with AI, but some like Python, R and Java are in vogue.

Generally, AI systems operate by consuming vast amounts of labeled training data, examining the data for patterns and correlations, and employing these patterns to predict future states. A Chabot fed with examples of text chats can thus learn to generate life like conversations with humans, or an image recognition system can learn to recognize and describe objects in images by looking at millions of examples.
AI programming is concerned with three cognitive abilities: learning, reasoning and self-correction.

Methodology

MODULE DESCRIPTION:

Data Pre-processing:

Machine learning validation techniques are employed to obtain the error rate of the Machine Learning (ML) model, which can be assumed to be near the actual error rate of the dataset. If the data volume is sufficiently large to represent the population, you might not require the validation techniques. But in real-time scenarios, to work with samples of data that are not necessarily a representative sample of population of given dataset. To obtaining the missing value, duplicate value and data type description whether it is float variable or integer.

Sample of data that has been used to give an unbiased estimate of a model fitted to training dataset for tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of dealing with content, quality, and structure of data can add up to a too-long to-do list. In the process of identification of data, it is helpful to know your data and its characteristics; this information will be useful in selecting what algorithm to use when constructing your model.

There are various types of data cleaning operations with Python's Pandas library and in particular, it addresses most likely the largest data cleaning operation, missing values and it can clean data more efficiently. It would like to have less time spent on cleaning data, and more time spent on exploring and modeling.

Some of these sources are mere random errors. Sometimes, there might be a more profound explanation for why data is missing. It is essential to learn these various forms of missing data from a statistical perspective. The form of missing data will affect how to approach filling in the missing values and to identify missing values, perform some simple imputation and elaborate statistical strategy for handling missing data. Prior to, joint into code, it is essential to learn the causes of missing data. Following are some common reasons for missing data:

User did not remember to enter a field.

Data was lost during manual transfer from a legacy database.

There was a coding mistake.

Users decided not to enter a field that was related to their assumptions regarding how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

import libraries for access and functional purpose and read the given dataset

General Attributes of Analyzing the provided dataset

Show the provided dataset in data frame form

show columns

shape of data frame

To define the data frame

Checking data type and dataset information

Checking for duplicate data

Checking Missing data values of data frame

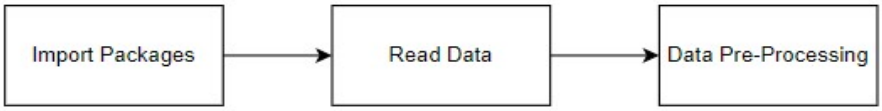
Checking unique data values of data frame

Checking count data values of data frame

Rename and drop the provided data frame

To define the type of values

MODULE DIAGRAM



Data visualization:

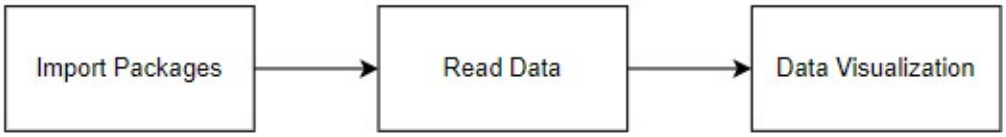
Data visualization is a valuable skill in applied statistics and machine learning. Statistics does indeed deal with quantitative descriptions and estimations of data. Data visualization offers an invaluable set of tools for acquiring a qualitative understanding. This can be useful when discovering and getting to know a dataset and can assist with detecting patterns, corrupted data, outliers, and much more. With a bit of domain knowledge, data visualizations can be employed to convey and illustrate central relationships in more visceral plots and charts than measures of association or significance. Data visualization and exploratory data analysis are entire disciplines in themselves and it will suggest a deeper look at some of the books listed at the end.

At times data is not understandable until it is able to view it in a graphical form, for instance with charts and plots. The ability to rapidly visualize of data samples and others is a very valuable skill in applied statistics and in applied machine learning. It will find the numerous forms of plots that you will be required to be familiar with when visualizing data in Python and how to utilize them to gain a better understanding of your own data.

How to plot time series data as line plots and categorical quantities as bar charts.

How to report data distributions in histograms and box plots.

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data

output : visualized data

Algorithm implementation:

Comparing the performance of various different machine learning algorithms regularly is valuable and it will find to create a test harness to compare several different machine learning algorithms in Python using scikit-learn. It can reuse this test harness as a starting point on your own machine learning problems and implement additional and varied algorithms to compare. Each model will have different performance profiles. You can use resampling techniques such as cross validation to obtain an estimate of how well each model will perform on unseen data. It must be able to use these estimates to select one or two top models from the range of models you've built. When we have a new dataset, it is wise to visualize the data in various ways so that we can examine the data from various angles. The same concept holds for model selection. You need to use several different methods of viewing the estimated accuracy of your machine learning models so that you can select the one or two to settle on. One method of doing this is to utilize various visualization techniques to display the average accuracy, variance and other characteristics of the distribution of accuracies of the models.

In the section that follows, you will find out how to do exactly that in Python with scikit-learn. The secret to making a reasonable comparison of machine learning algorithms is that each algorithm has to be compared in the same manner on the same data and it can make it do this by compelling every algorithm to be compared on an identical test harness.

Performance Metrics to calculate

False Positives (FP): An individual who will pay predicted as defaulter. When actual class is no and predicted class is yes. For example, if actual class says that this passenger didn't survive but predicted class says that this passenger will survive.

False Negatives (FN): An individual who default predicted to be payer. When actual class is yes but predicted class is no. Example: if actual class value reveals that this passenger survived and predicted class informs you that passenger will die.

True Positives (TP): An individual who will not pay anticipated as defaulter. These are the correctly anticipated positive values which implies that the value of actual class is yes and value of predicted class is also yes. For example, if actual class value is showing that this passenger survived and predicted class informs you the same.

True Negatives (TN): An individual who default predicted as payer. These are the correctly predicted negative values i.e. the value of actual class is no and value of predicted class is also no. For example if actual class tells you that this passenger did not survive and predicted class also says the same thing.

True Positive Rate(TPR) = TP / (TP + FN)
False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Ratio of the total number of predictions which is correct otherwise in general how frequently the model predicts correctly non-defaulters and defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most natural measure of performance and it is merely a ratio of correct predicted observation to total observations. You might be thinking that, if we have high accuracy then our model is optimal. Yes, accuracy is good but only when you have symmetric datasets where values of false positive and false negatives are nearly equal.

Precision: The ratio of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the number of positive observations predicted correctly divided by the number of total positive observations predicted. The question that this measure answer is of all passengers who were labelled as survived, how many survived? High precision is associated with low false positive rate. We have achieved 0.788 precision which is quite good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will predict accurately)

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the harmonic mean of Precision and Recall. Hence, this score considers false positives and false negatives as well. Instinctively, it's less easy to get one's head round than accuracy, but F1 will tend to be more informative than accuracy if you've got an imbalance between the classes. Accuracy will be optimal only when false negatives and false positives incur comparable cost. When false positive and false negative cost greatly differs, you would do well to consider Precision as well as Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

The below 4 different algorithms are compared:

MLP Classifier Algorithm

Bagging Classifier Algorithm

Linear Regression Algorithm

Random Forest Regression

Conclusion and Future work

In summary, crop yield forecasting with machine learning exploits sophisticated algorithms to process huge volumes of farm data, allowing for precise forecasted yields. This practice assists farmers in making rational decisions, improving the utilization of resources, and enhancing farm productivity. Through the utilization of real-time data and predictive analytics, machine learning systems are capable of greatly streamlining crop farming practices and enabling sustainable farming.

Spreading the project on the cloud.

To optimize the work to implement in the IOT system.

References

TITLE: CROP YIELD PREDICTION USING MACHINE LEARNING

AUTHOR: Sreenidhi Chappidi

YEAR: 2022

TITLE: CROP YIELD PREDICTION USING MACHINE LEARNING

AUTHOR: Mayank Champaneri, Chaitanya Chandvidkar

YEAR: 2017

TITLE: Exploring the suitability of machine learning algorithms for crop yield

AUTHOR: M. VARMA, K. N. SINGH

YEAR: 2022

TITLE: CROP YIELD PREDICTION USING MACHINE LEARNING

AUTHOR: Ishwarya, Nagapooja BN