Performance Evaluation of Predictive Models for Missing Data Imputation in Weather Data

Doreswamy
Department of Computer Science,
Mangalore University,
Karnataka, INDIA
Email: doreswamyh@yahoo.com

Ibrahim Gad
Department of Computer Science,
Mangalore University,
Karnataka, INDIA
Faculty of Science, Tanta University,
EGYPT
Email: gad_12006@yahoo.com

B.R. Manjunatha
Department of Marine Geology,
Mangalore University,
Karnataka, INDIA
Email: omsrbmanju@yahoo.com

***Abstract*—Real datasets can have missing values for a different reasons such as in data that were not kept on file and data corruption. Climate forecasting has a highly relevant effect in agricultural fields and industries sectors. The process of predicting climate conditions is required for different areas of life sectors. Handling missing data is significant because a lot of machine learning algorithms performance are affected by missing values in addition, they do not support data with missing values. Various techniques have been used to process missing data problem and the most applied is removing any row that contains at least one missing value. Also, another approaches to solve missing data problems are to impute the missing data to yield a more complete dataset. In order to improve the accuracy of prediction with the climate data, missing value from dataset should be removed or imputed/predicted in the pre-processing phase before using the data for prediction or clustering in the analysis step. In this paper, we propose a new technique to handle missing values in weather data using machine learning algorithms by execute experiments with NCDC dataset to evaluate the prediction error of five methods namely the kernel ridge, linear regression, random forest, SVM imputation and KNN imputation procedure. The missing values were imputed using each method and compared to the observed value. Results of the proposed method were compared with existing techniques.**

***Keywords*—Missing data, Imputation, NCDC data set, Weather analysis, SVM, KNN.**

I. Introduction

The most common problem in real dataset and statistical analysis is missing data. The percentage of missing values varies from one dataset to another. Generally, the dataset contains different percentages of missing values in each column [20]. Usually, if the rates of missing data is less than 1% are called trivial and missing data ratio with in the range of $1-5\%$ is flexible. The advanced methods are applied to handle rates with in the range $5-15\%$, and greater than 15% have a very great impact on analysis [4, 12].

The process of missing data estimation is an essential problem in data analysis , and there are several solutions have been suggested to solve such problem , for instance, statistics and data mining [16, 17]. The most familiar approaches to handle missing values involves removing all instances with missing values from a dataset and imputing missing values [22, 25]. Imputation is a method to assess the missing data based on the complete instances in a dataset. The types of imputation methods are parametric and nonparametric regression imputation methods [11]. In the imputation strategy such as data mining and machine learning algorithms, missing data handling is independent of data analysis algorithms. Usually, the observed data in incomplete rows is used to estimate missing values by imputation algorithms. For example, KNN imputation algorithm depends on the available data to indicate neighbours of an instance with missing values , and the class of the instance in clustering-based imputation algorithms [8, 24].

There are three categories of treatment methods for missing data namely :- a) case deletion, which is the most common used. So that each row containing missing values is deleted. b) parameter estimation can be done by maximum likelihood procedures that are used to treat parameter estimation. In general, the methods of parameter estimation are more efficient than case deletion methods, because they can use all the data available in the dataset. However, parameter estimation suffers from the following limitations: a high sensitivity to outliers, and a high degree of complexity and c) imputation techniques: the definition of imputation is a procedure to fill missing data with predicted ones based on values available in the dataset [6, 10].

In this paper the following techniques are selected to fill missing data: the kernel ridge , linear regression, random forest, SVM imputation and the k-nearest neighbor (KNN) imputation procedure. In a NCDC dataset, the attributes that have missing values are related with temperature attribute then by using known values of temperature we can predict the unknown values of the other attributes. For evaluating the results, the performance metrics considered are standard deviation of error (STDE), variance of error (VARE), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), bias and coefficients of determination R^2. The remaining sections of this paper is organized as follows. Section II shows the related works. Section III describes the NCDC dataset. Section IV describes the five selected methods to handle missing values in this paper. Section V explains the steps of the proposed model. Section VI presents the criteria that used to compare among the selected methods. Section VII presents and discuss comparison results. Finally, conclusion is presented in Section VIII.

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

II. Related Work

Swati et all [10] proposed a methodology consists of two phases. The first phase is missing value check and outlier detection, this step is pre-processed step of missing data and missing value locations are checked in the input dataset. The second step is calculating estimation of missing values, firstly small array is created from the input data in which missing data value is existing. secondly, calculate centroid of the subset, centroid is generated by the mean of subset. The value of X_{est}(estimated value) is separately computed for every missing value in the complete dataset.

Gimpy et all [21] proposed estimation of missing values using decision tree approach. They explored that the performance of classifier affected by the existence of missing values in a dataset. Dataset that used in this study contains some missing values, which consists of student data of university system. Classification algorithm (C4.5/J48) is used to fill missing values and the accuracy is measured by confusion matrix.

Engels et all [9] suggested a fourteen types of mean imputation techniques introduced on missing data(Column mean - Column median - Class mean - Class median - Hot deck - Regression - Regression with error - Previous row mean - Previous row median - Last observation carried forward - Row mean - Row median - Next observation carried backward - Average of the last known and next known values). The missing value was imputed using each method and compared to the observed value. Methods were compared on the root mean square error, bias, mean absolute deviation and relative variance of the estimates.

Sallam et all [18] suggested handling numerical missing values using rough sets. They investigate multiple ways used to solve the problem of missing values in a a dataset. The proposed model used rough set theory as a technique to fill missing data. In addition, the model has ability to estimate the missing values for condition and decision attributes.

Patidar et all [15] proposed handling missing value using decision tree algorithms. They apply data mining techniques to analysis the performance of the students in educational system. C5.0 decision tree is used to make analysis and making decisions. Also, comparison is accomplished by ID3, C4.5 and C5.0. This research explored that C5.0 gives more accurate and efficient output than the other methods.

III. NCDC Dataset Description

This section explains the details of the data of day product produced by the National Climatic Data Center (NCDC) dataset with missing values. National Climatic Data Center (NCDC) has more than 9000 stations data available online [1, 2]. As shown in Figure 1, we have selected 79 stations having 16 columns(variables) and 365 rows from NCDC dataset for our work for the years from 2000 to 2016. The variable names are described in Table I.

NCDC dataset is known to have a lot of missing values. Specifically, there are missing observations for some columns that are marked as 9999.9, 999.9 or 99.99. We can confirm this by the definition of those columns that a 9999.9, 999.9 or 99.99 value is invalid for those measures, for instance, a 9999.9 indicates missing value for the columns Mean temperature (TEMP), Mean dew point (DEWP), Mean sea level pressure (SLP), Mean station pressure (STP), Maximum temperature (MAX) and Minimum temperature (MIN). The value 999.9 indicates missing value for the columns Mean visibility (VISIB), Mean wind speed (WDSP), Maximum sustained wind speed (MXSPD), Maximum wind gust (GUST) and Snow depth (SNDP). Moreover, the value 99.99 indicates missing value for the columns Total precipitation (PRCP). Table II shows sample of the selected stations of India and the missing values for each attribute of the NCDC dataset.

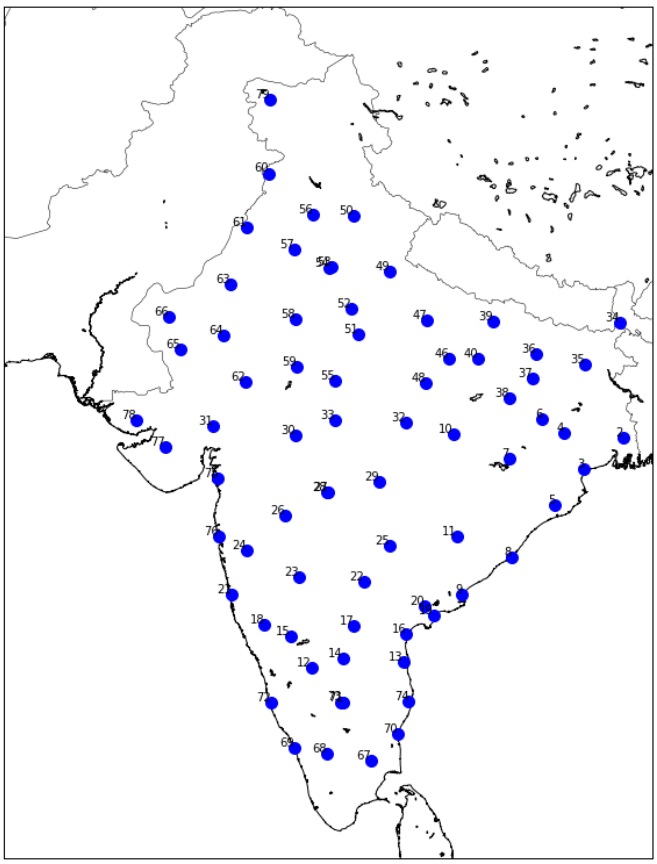

Fig. 1. The selected 79 stations of India from NCDC dataset.

TABLE I. Description of the attributes for each station.

No.	Attribute	Description
1	STN	Station number
2	WBAN	Weather Bureau Air force Navy
3	YEARMODA	The year, month and day
4	TEMP	Mean temperature
5	DEWP	Mean dew point
6	SLP	Mean sea level pressure in millibars
7	STP	Mean station pressure
8	VISIB	Visibility
9	WDSP	Mean wind speed in knots
10	MXSPD	Maximum sustained wind speed in knots
11	GUST	Maximum wind gust
12	MAX	Maximum temperature
13	MIN	Minimum temperature
14	PRCP	Precipitation
15	SNDP	Snow depth in inches
16	FRSHTT	Fog rain snow hail thunder Tornado

TABLE II. NUMBER OF MISSING VALUES IN ATTRIBUTES OF INDIAN STATIONS OF NCDC DATASET.

STN	TEMP	DEWP	SLP	STP	VISIB	WDSP	MXSPD	GUST	MAX	MIN	PRCP	SNDP
4202700	0	6	6017	18	3	2	64	6017	0	1	635	6011
4207100	0	9	4701	6158	16	20	124	5938	0	0	253	6157
4210100	0	2	1	5784	0	2	453	5784	0	1	131	5782
4211100	0	5	8	5	0	0	40	6141	0	0	169	6139
4212300	0	28	35	3926	8	13	583	3926	0	0	203	3926
4213100	0	0	2	6155	0	0	121	6155	0	0	282	6154
4216500	0	3	3	5638	0	1	74	5638	0	0	49	5638
4218100	0	0	6088	6091	0	1	5	4948	0	7	944	6091
4218200	0	0	6	5828	0	0	118	5804	0	0	127	5828
4218900	0	11	15	5716	1	5	549	5716	0	0	126	5716

IV. DIFFERENT ALGORITHMS TO IMPUTE MISSING VALUES

This section briefs description of the predictive algorithms to treat missing values.

A. Linear Regression

Linear Regression is a simple and classic method that fits a line through the dataset that consists of a set of features [13, 19]. The general mathematical form of linear regression is:

$$y_i = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_i x_i \cdots + w_m x_m + \varepsilon_i \quad (1)$$

where: y_i is the i^{th} observation of the dependent variable, x_i is the i^{th} observation of the independent variable, w_0 is intercept term, w_i is slope coefficient for i^{th} independent variable, ε_i is the error term for the i^{th} observation and m is the number of independent variables.

Equation 2 shows mathematical form for a feature space of m features

$$f_w(x) = \sum_{j=0}^{m} x_j w_j + \varepsilon \quad (2)$$

As shown in Equation 3, we need to minimize the value of error term so as to get the optimal solution for the predictive purpose. There are several approaches for finding the optimal values of weights.

$$Error(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 \quad (3)$$

A gradient descent is a numerical method commonly used to solve this problem. Linear regression algorithm has the following advantages: it is a common method and it has an intuitive interpretation. Based on the magnitude of the weights, it will tell which parameters are the most influence and if there is a positive or negative correlation.

B. K Nearest Neighbours Imputation (KNNI).

In KNN, to predict the missing values it uses average of similar instances to the interested instance. The similarity of two instances calculated by distance function. Minkowski distance function which is available in scikit-learn is used to compute the similarity of instances [23].

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (4)$$

By default p=2 and thus this reduces to the Euclidean distance.

Pseudo-code 1 shows the steps of KNN algorithm:

Algorithm 1 KNN Algorithm [4]

Require: Divide the data set D into two parts. Let D_m be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called D_c and object $z \in D_c$.

1: **Begin**
2: For each vector x in D_m:
a) Divide the instance vector into observed and missing parts as $x = [x_o; x_m]$.
b) Compute $d(x_o, z)$, the distance between the x_o and z. Use only those features in z, which are observed in the vector x.
c) Select the K closest instances vectors (K-nearest neighbors) to x
d) Replace the missing value using the mean value of the attribute in the K-nearest neighborhood.
3: **End**

KNN considers a different approaches. The core assumption made with this method is that examples with similar feature vectors should have similar outputs. Neighbours of a queried data point should be used to determine its value and not points that are far away. In case of a regression, the values of k neighbours are selected and their mean is considered to set the value of the queried point.

C. Imputation using a prediction model

The basic idea of this method depends on predictive model constructed to predict the values that will replace the missing data. Generally, the column with 0% of missing data is the independent variable and the remaining columns are dependent variables for the model used. The drawbacks of this approaches are (i) If any attribute has missing data does not have relationship among the other attributes in the dataset then the performance of model will be worst (iii) to construct a large number of predictive models to predict missing values is highly computational cost [4].

D. Decision tree algorithms

There are various algorithms of decision tree that are used to solve the problem of missing values by applying different built-in approaches. For instance, the basic idea of C.4.5 depends on probabilistic approach in order to treat missing data in training and testing dataset, respectively. Besides, CART method uses the values of surrogate attribute, which has the

strongest correlation with the given attribute to fill missing values. To increase the predictive performance and solve over-fitting problem, the random forest algorithm applies a number of decision tree methods on number of samples selected from the dataset [15, 21, 23].

E. Support vector machine (SVM)

SVM algorithm is a supervised learning used for different purpose such as regression, classification and outliers detection. The basic idea of support vector machine depends on constructing a set of hyper-planes in a space, and the hyper-plane that has the longest distance to the closet training data points of any class is considered a good separation line. Support vector regression (SVR) is a type of support vector classification has ability to handle regression problems. The support vector regression model depends only on a subset of the training points, because the cost function for building the model ignores any point close to the model prediction [5, 23].

Equation 5 describes the mathematical form of support vector regression. $f(x)$ is the prediction function, ε-insensitive loss function, C is a constant and w is the slope of line. Equation 6 shows the input data for the model, the model complexity term $||w||^2$ shown in Equation 7, and the regularized risk functional described in Equation 7.

$$f(x) = (w.x) + b, \qquad w, x \in \mathbb{R}^N, b \in \mathbb{R} \tag{5}$$

input data for the model

$$(x_1, y_1), \cdots, (x_i, y_i) \in \mathbb{R}^N \times \mathbb{R} \tag{6}$$

$$||w||^2/2 + C \cdot \mathbb{R}^{\varepsilon}_{emp} \tag{7}$$

$$\mathbb{R}^{\varepsilon}_{emp} = \frac{1}{l} \sum_{i=1}^{l} |y_i - f(X_i)|_{\varepsilon} \tag{8}$$

$$|y - f(X)|_{\varepsilon} = max\{0, |y - f(X)| - \varepsilon\} \tag{9}$$

The kernel ridge regression (KRR) model is constructed by combining two methods namely the kernel trick and Ridge Regression that consists of linear least squares with l2-norm regularization. The kernel ridge model learns a linear function in the space that constructed by kernel and data. So, for non-linear function in the original space there exist corresponding each non-linear kernels. The Kernel Ridge and support vector regression (SVR) have the same learning form of the model. But, they different in loss functions are used, KRR uses squared error loss while SVR uses ϵ-insensitive loss, both combined with l2 regularization [14].

V. The proposed Method for Missing data imputation

In this section, we explain how to identify and mark values as missing in NCDC dataset. The process of identify missing or corrupt data executed by plots and summary statistics. The dataset is loaded as a Pandas Data Frame and print summary statistics on each attribute for each weather data station of India. Experiments were carried out using 79 of stations in India downloaded from the NCDC website [1, 2]. Table II presented a summary of the common characteristics of each station in the dataset. We can see that there are columns that have a lot of missing values. On some columns, a value of 9999.9, 999.9 or 99.99 does not make sense and indicates an invalid or missing value. Specifically, the following columns of station 4202700 have 9999.9, 999.9 or 99.99 an invalid value. The following columns SLP, GUST and SNDP have the highest percentage of missing values consequently they are not imputed here.

Figure 2 presented the proposed model to estimate the missing values in the selected stations. First, we selected 79 stations having a different percentage of missing data in each attribute. Each station passed through a cleaning phase where attributes with more than 50% of missing value were eliminated. The goal of cleaning step is to reduce the number of imputations in order to increase the accuracy of missing value prediction. After cleaning processes, we select the attribute which has the lowest number of missing value to impute the missing value for that attribute, for instance MIN column has the lowest number of missing values. So, the first column selected is MIN. The original dataset is divided into two datasets namely Full_df and Missing_df such that the indexes of full rows of MIN column is indicated for Full_df dataset and the indexes of missing values of MIN column for Missing_df dataset. Next, Full_df is divided into Train and Test datsets in order to train and test the machine learning algorithms. Then, we apply the five methods to treat missing values. in addition, compute estimates of the error1 (MSE, RMSE, VARE, STDE, MAE, R^2, BIAS) by 5-fold cross-validation. Missing_df is used as input for machine learning algorithm to predict the missing values of MIN column and fill it.

Second, the station dataset rows without missing values and the complete version of original dataset after filling the missing values are merged to get final dataset without missing data. The same steps are repeated again, that is, after filling the missing values, the original dataset is divided into Train and Test datsets in order to train and test machine learning algorithms. Then, the proposed model was train and tested using the 5-fold cross-validation estimates of the error2 (MSE, RMSE, VARE, STDE, MAE, R^2, BIAS).

Finally, after missing values of MIN attribute imputed we select the next attribute among the remaining attributes to impute which has the smallest number of missing data after MIN column. To predict the missing values of the incomplete valued attributes, we consider the TEMP attribute as independent variable since it doesn't have missing values, as shown in Table III.

VI. Comparison of methods

The proposed framework has been experimented with 79 stations dataset. A detailed explanation about the results obtained during the imputation process using the proposed framework is discussed in this section. There are different approaches to calculate the performance of predictive models [3]:

TABLE III. THE ORDER OF ATTRIBUTES DURING THE PROCESS OF IMPUTATION.

Independent variable	Dependent variable	Number of missing values
TEMP	MIN	1
TEMP	WDSP	2
TEMP	VISIB	3
TEMP	DEWP	6
TEMP	STP	18
TEMP	MXSPD	64
TEMP	PRCP	635
TEMP	SNDP	6011
TEMP	SLP	6017
TEMP	GUST	6017

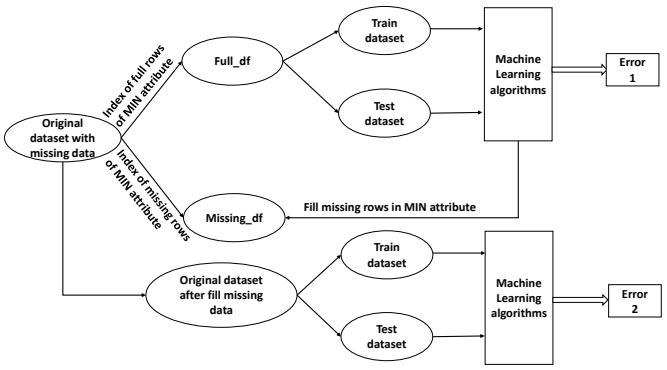

Fig. 2. The prediction model to impute missing values.

- Estimator score method: is used to provide evaluation criterion that is different from one problem to another.
- Scoring parameter: In this approach the model-evaluation is done by using cross-validation which depend on an internal scoring strategy.
- Metric functions: Evaluation metrics is used to interpret the performance of a model. The importance of evaluation metrics is their ability to differentiate among model results by metric functions, which are used to evaluate prediction error of the model. For example: Regression metrics are Mean absolute error, Mean squared error, and the coefficient of determination R^2 .

After imputing the missing data, the following summary measures is used to calculate the prediction performance. Let y_i be the i^{th} observed value, $\hat{y}_i$ be the i^{th} estimated value, $\bar{y}$ be the mean of observed values and n is the number of observed values. Equations (10) – (15) define the MAE, RMSE, MSE, bias and R^2 measures, respectively.

A. Mean Absolute Error (MAE)

MAE is a measure that represents the positive and negative deviations between the predicted and the observed values [7]. Mathematically, it is defined as:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| . \tag{10}$$

The Mean Absolute Deviation (MAD) and the Mean Absolute Error (MAE) are the same method for measuring prediction error [7]. Mathematically, MAD is defined as:

$$MAD(y, \hat{y}) = \frac{\sum_{i=1}^{n} |y - \hat{y}|}{n} \tag{11}$$

B. Mean Squared Error (MSE)

MSE measures the accuracy of machine learning algorithms. Mathematically, it is defined as:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{12}$$

C. Root Mean Squared Error (RMSE)

The root-mean-square error (RMSE) or root-mean-square deviation (RMSD) measures the average squares of the errors [7]. Mathematically, it is defined as:

$$RMSE = \sqrt{\frac{\sum(y - \hat{y})^2}{n}} \tag{13}$$

An additional summary measure is bias, defined as:

$$BIAS = \frac{\sum_{i=1}^{n}(y - \hat{y})}{n} \tag{14}$$

D. R^2 score, the coefficient of determination

The R^2, the coefficient of determination, is a number that provides a measure of correlation between the observed and estimated values. It varies between 0 and 1. The highest possible score is 1.0 which indicates strong correlation, a value of 0 indicates no agreement and it can be negative [3]. Mathematically, it is represented as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

VII. RESULTS

To decrease the effect of missing values on the prediction precision, we selected only variables that have percentage of missing values less than 50% in each station. The results of different analyses are noted briefly in the following subsections.

A. Descriptives

Table IV summarizes the dataset in terms of the number of dependent attributes identified for each station's dataset. Table IV gives descriptive statistics for station number(STN 4202700) and the missing data rate in the years 2000 to 2016. As can be seen, the missing data rate is different from one attribute to another across stations. The number 6017 in column SLP indicates the number of missing values for station 4202700. There are 6 missing values in DEWP column among 6017 instances of station number(STN 4202700) in the period of time between 2000 and 2016. The value 0.099 represents the percent of the DEWP instances were missed before the missing values were imputed. MIN (1), WDSP (2), VISIB

TABLE IV. DESCRIPTIVE STATISTICS AND MISSING DATA RATE FOR STATION NUMBER(STN 4202700).

	TEMP	DEWP	SLP	STP	VISIB	WDSP	MXSPD	GUST	MAX	MIN	PRCP	SNDP
count	6017	6011	0.0	5999	6014	6015	5953	0.0	6017	6016	5382	6
mean	56.53	45.77	NaN	840.89	2.19	0.94	2.43	NaN	69.69	46.03	0.044	4.26
std	14.21	12.43	NaN	4.91	0.70	0.68	3.39	NaN	15.18	13.46	0.18	2.07
min	27.50	17.50	NaN	820.50	0.10	0.00	1.00	NaN	30.20	15.30	0.00	1.60
25%	43.20	34.80	NaN	837.20	1.90	0.60	1.0	NaN	56.30	34.20	0.00	2.57
50%	57.30	45.30	NaN	841.50	2.30	0.80	1.90	NaN	71.60	45.90	0.00	4.70
75%	69.30	56.50	NaN	844.60	2.50	1.10	2.90	NaN	83.10	57.60	0.0	5.70
max	86.80	71.80	NaN	868.10	11.30	12.70	49.90	NaN	102.60	75.90	2.48	6.70
missing	0	6	6017	18	3	2	64	6017	0	1	635	6011
% missing	0	0.099	100	0.29	0.049	0.033	1.06	100	0	0.016	10.55	99.90

(3) columns contains small percentages of missing value. The average missing value of (GUST , SLP , SNDP) is about 99% points or higher than the mean for non-missing values for the same attributes, SLP (6017), GUST (6017), SNDP (6011) columns contain the highest percentage of missing values, respectively. On the contrary, the TEMP and MAX columns have the smallest percentage of missing values for station 4202700 dataset. TEMP and MAX columns are 0 points hence, there was no missing values. The following columns of station 4202700 SLP, GUST and SNDP were neglected during imputation process because of they have high number of missing values.

B. Accuracy results

To measure the performance of the proposed model, we compute 5-fold cross-validation estimates of the prediction error for all methods. The results are shown in Table V.

The RMSE is used to measure the performance of the different predictive methods that treat the missing values. Table V shows the average value (calculated from the 5 - fold cross-validation) of the RMSE. Obviously, Random Forest and Linear Regression method had the smallest mean RMSE among all methods, MIN (0.07) and PRCP (0.1839), respectively. Performance gradually became worse for the other methods. Table V shows that KNN, SVM, and Kernel ridge had substantially worse accuracy than the other methods, with an average RMSE greater than 4, DEWP (4.06), MIN (4.21), and DEWP (4.17), respectively. The MSE, VARE, MAE, STDE have the same results like RMSE where Random Forest and Linear Regression methods had the smallest mean error among all methods. In addition, KNN, SVM, and Kernel ridge had worse accuracy than the other methods.

C. Bias results

Table V shows the 5-fold cross-validation error rates for all predtictive algorithms. The mean deviation (MD) was used to assess the bias, with a MD of zero indicating no bias. A positive bias indicates that on average, the imputed value underestimated the true value. Table V shows that linear regression method had little bias for DEWP, MIN, PRCP and STP columns. Mean biases for this method was DEWP(0.000288), MIN(0.000737), PRCP(0.000017) and STP(0.000055), respectively. The highest biased methods were the SVM for WDSP(0.118) column and Kernel Ridge for STP(0.287).

D. R^2 results

Table V shows that KNN, SVM, and Random Forest had substantially worse R^2 value than the other methods, with an average R^2 less than 0 for MXSPD, PRCP, STP, VISIB and WDSP. The Kernel ridge and Linear Regression method had the highest mean R^2 among all methods, MIN (0.91) and DEWP (0.89), respectively. Performance gradually became worse for the other methods. Also, some values of R^2 are negative values. Because the models are not fitted well and the relationship between independent and dependent variables is negative. So, it is harder to predict the missing value , Figure 3.

In addition, Figure 3 summarizes the correlation between dependent and independent attributes such that, the X-axis in Figure 3 denotes the dependent attributes that will be considered for each station. The Y-axis in Figure 3 denotes its corresponding independent attribute.

Figure 3 indicates the correlation between Temp column and the rest of the other columns. It is clearly shown that figure Figure 3 (a) and (b) has strong correlation with temperature as they have positive slope. The rest of the other columns have weak correlation which is indicated by small values or negative correlation which is shown by negative value, as shown in Figure 3 (c), (d), (e), (f) and (g), respectively. Table VI shows comparison between different algorithms for error 1 and error 2 of the proposed model.

VIII. CONCLUSION

The main contribution of this work is to predict missing values from known data using machine learning techniques on weather data. In this work, the performance of proposed methods is more reliable according to the comparison among the selected imputation techniques. Comparing the R^2 we can see that almost all methods performed efficiently in DEWP and MIN columns, since the percentages of the missing values is low. Overall, all imputation methods seem to be performing better when the percentage of missing values increase and the correlation between independent and dependent variables is positive strong. The procedures of replacing missing data with a value of zero or the same value can have an adverse effect by generating outliers and noise in the data. In addition, to remove attributes that contain missing data as this will negatively affected the size of dataset and performance of prediction model. The results generally were not as great as was expected. The future direction is to improve the performance of the proposed algorithms so as to use for different other weather stations across India.

TABLE V. COMPARISON OF IMPUTATION RESULTS OF ALL PREDICTIVE METHODS FOR NCDC DATA.

Performance Measure	Column Name	Kernel Ridge	Linear Regression	Random Forest	SVM	KNN
RMSE	DEWP	4.178127	4.008448	0.083062	4.144169	4.061828
	MIN	4.163366	3.832284	0.074882	4.218798	3.875922
	MXSPD	3.372359	3.386554	0.080938	3.420199	3.471447
	PRCP	0.18444	0.183925	0.08723	0.191566	0.189368
	STP	3.02673	3.20979	0.154479	3.005521	3.031456
	VISIB	0.631192	0.639718	0.096721	0.634688	0.650434
	WDSP	0.674293	0.67517	0.084378	0.68755	0.692234
MSE	DEWP	17.472703	16.073027	0.006909	17.198758	16.507689
	MIN	17.356274	14.688889	0.005628	17.842198	15.028945
	MXSPD	11.511962	11.496899	0.006702	11.938489	12.182443
	PRCP	0.03435	0.03434	0.007649	0.037508	0.03617
	STP	9.171538	10.310928	0.029123	9.062258	9.198546
	VISIB	0.400172	0.410611	0.009596	0.404029	0.424362
	WDSP	0.461462	0.461661	0.007678	0.473852	0.485024
VARE	DEWP	16.09126	16.051852	0.006569	16.320702	16.492945
	MIN	15.025009	14.670625	0.004779	15.504685	15.015872
	MXSPD	11.488536	11.49572	0.006665	11.493234	12.17077
	PRCP	0.034301	0.034257	0.007626	0.034446	0.036141
	STP	8.920676	10.301495	0.009565	8.894697	9.18946
	VISIB	0.399323	0.410468	0.005636	0.402297	0.423869
	WDSP	0.460357	0.461181	0.005917	0.45954	0.48457
MAE	DEWP	3.249414	3.118716	0.064831	3.229003	3.181393
	MIN	3.41079	3.171125	0.061617	3.47117	3.180699
	MXSPD	1.463751	1.462091	0.034429	1.241584	1.548302
	PRCP	0.075231	0.075713	0.032937	0.119918	0.076576
	STP	2.264162	2.448606	0.1331	2.234247	2.25061
	VISIB	0.38647	0.409582	0.073045	0.389695	0.413837
	WDSP	0.408057	0.408798	0.051983	0.393936	0.423598
STDE	DEWP	4.009941	4.005804	0.08102	4.037656	4.060013
	MIN	3.874821	3.829906	0.069117	3.934757	3.87423
	MXSPD	3.368322	3.38637	0.080743	3.355377	3.469782
	PRCP	0.184314	0.183694	0.087106	0.182905	0.189289
	STP	2.985573	3.208272	0.095903	2.978754	3.029975
	VISIB	0.630503	0.639605	0.074737	0.633373	0.650055
	WDSP	0.673484	0.674819	0.075634	0.677141	0.691905
BIAS	DEWP	-0.509053	0.000288	-0.014315	-0.651159	0.017721
	MIN	-0.421798	0.000737	-0.017999	-0.618795	0.007572
	MXSPD	-0.003881	-0.000023	0.003814	0.659695	0.003464
	PRCP	0.000704	0.000017	0.003613	-0.054969	-0.000314
	STP	0.287231	0.000055	0.059896	0.058289	-0.004909
	VISIB	-0.005467	-0.000042	0.031274	0.037779	-0.001092
	WDSP	-0.003153	-0.000005	0.028695	0.118175	0.001526
R^2	DEWP	0.886933	0.895785	0.882179	0.888191	0.893024
	MIN	0.90398	0.918842	0.902151	0.901528	0.916975
	MXSPD	-0.000467	0.0019	-0.045134	-0.037314	-0.0618
	PRCP	0.003244	0.002607	-0.033585	-0.106879	-0.052167
	STP	0.619772	0.572162	-0.290697	0.624228	0.61834
	VISIB	0.19289	0.173096	-0.429562	0.18586	0.143497
	WDSP	0.026231	0.024218	-0.210486	-0.001842	-0.026706

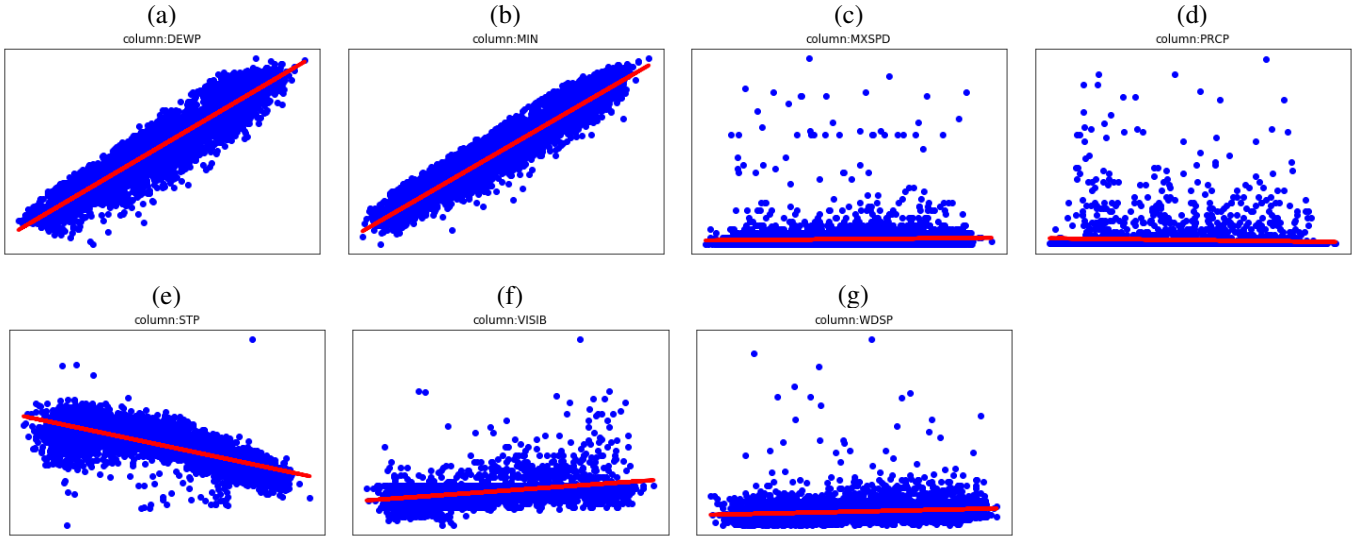

Fig. 3. The correlation between Temp column and the others columns (a)DEWP, (b)MIN, (c)MXSPD, (d)PRCP, (e)STP, (f)VISIB and (g)WDSP, respectively.

TABLE VI. COMPARISON BETWEEN DIFFERENT ALGORITHMS FOR ERROR 1 AND ERROR 2 OF THE PROPOSED MODEL.

Performance Measure	Column Name	Kernel Ridge		Linear Regression		Random Forest		SVM		KNN	
		Error 1	Error 2	Error 1	Error 2	Error 1	Error 2	Error 1	Error 2	Error 1	Error 2
RMSE	DEWP	4.178127	0.775016	4.008448	0.777233	0.083062	0.786469	4.144169	0.961779	4.061828	4.092769
	MIN	4.163366	4.298482	3.832284	4.037982	0.074882	4.323557	4.218798	4.254737	3.875922	3.893093
	MXSPD	3.372359	3.798867	3.386554	3.795462	0.080938	3.882461	3.420199	3.907722	3.471447	3.800336
	PRCP	0.18444	0.184435	0.183925	0.183375	0.08723	0.190140	0.191566	0.193106	0.189368	0.189077
	STP	3.02673	3.710909	3.20979	3.699253	0.154479	3.796066	3.005521	3.880287	3.031456	2.978395
	VISIB	0.631192	0.615462	0.639718	0.616140	0.096721	0.631542	0.634688	0.660633	0.650434	0.635328
	WDSP	0.674293	41.918055	0.67517	3.132149	0.084378	43.534611	0.68755	3.235354	0.692234	0.790139

ACKNOWLEDGMENT

The authors thank for financial assistance provided by the SERB-DST vide SB/EMEQ-137/2014, dated 21-03-2016, the ICCR fellowship by the Governments of India, and Egypt for carrying the project.

REFERENCES

[1] Ncdc climate services, . URL ftp://ftp.ncdc.noaa.gov/pub/data/gsod.

[2] Ncdc climate services, . URL https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets.

[3] scikit-learn, machine learning in python. URL http://scikit-learn.org/stable/modules/model_evaluation.html.

[4] E. Acuña and C. Rodriguez. The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, pages 639–647, 2004.

[5] I. B. Aydilek and A. Arslan. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233:25–35, 2013.

[6] E. C. Blessie, E. Karthikeyan, and V. Thavavel. An extended relief-disc for handling of incomplete data to improve the classifier performance. In *Mathematical Modelling and Scientific Computation*, pages 530–536. Springer, 2012.

[7] L. Campozano, E. Sánchez, A. Aviles, and E. Samaniego. Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the ecuadorian andes. *Maskana*, 5(1):99–115, 2015.

[8] J. Chen and J. Shao. Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453):260–269, 2001.

[9] J. M. Engels and P. Diehr. Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, 56(10):968–976, 2003.

[10] S. Jain and K. Jain. Estimation of missing attribute value in time series database in data mining. *Global Journal of Computer Science and Technology*, 16(5), 2017.

[11] R. J. Little and D. B. Rubin. *" Statistical analysis with missing data "*. John Wiley & Sons, 2014.

[12] J. Luengo, S. García, and F. Herrera. A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfns and eventcovering method. *Neural Networks*, 23(3):406–418, 2010.

[13] D. C. Montgomery, E. A. Peck, and G. G. Vining. *" Introduction to linear regression analysis "*. John Wiley & Sons, 2015.

[14] K. P. Murphy. *" Machine learning: a probabilistic perspective "*. MIT press, 2012.

[15] P. Patidar and A. Tiwari. Handling missing value in decision tree algorithm. *International Journal of Computer Applications*, 70(13), 2013.

[16] R. K. Pearson. *" Mining imperfect data: Dealing with contamination and incomplete records "*. SIAM, 2005.

[17] M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.

[18] E. Sallam, T. Medhat, A. Ghanem, and M. Ali. Handling numerical missing values via rough sets. *International Journal of Mathematical Sciences and Computing(IJMSC)*, 3(2):22–36, 2017.

[19] G. A. Seber and A. J. Lee. *" Linear regression analysis "*, volume 936. John Wiley & Sons, 2012.

[20] L. Sunitha, M. BalRaju, and J. Sasikiran. Data mining: Estimation of missing values using lagrange interpolation technique. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, 2: 1579–1582, 2013.

[21] G. Vohra, Rajan and Minakshi. Estimation of missing values using decision tree approach. *(IJCSIT) International Journal of Computer Science and Information Technologies*, 5(4):5216–5220, 2014.

[22] Q. Wang, J. Rao, et al. Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, 30(3):896–924, 2002.

[23] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg.

[24] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang. Missing value imputation based on data clustering. In *Transactions on computational science I*, pages 128–138. Springer, 2008.

[25] S. Zhang, Z. Jin, and X. Zhu. Missing data imputation by utilizing information within incomplete instances. *Journal of Systems and Software*, 84(3):452–459, 2011.