Predicting Superconducting Critical Temperatures with Uncertainty Quantification using Supervised Machine Learning

K. Kleinasser, Cornell University, Ithaca, NY*

Abstract will be added here.

I. INTRODUCTION

Superconductors are materials that lose all electrical resistance at low temperatures. These materials have a critical temperature (T_C) at which they lose their resistance. Most have very low critical temperatures, but "unconventional superconductors" can have critical temperatures as high as room temperature under non-atmospheric conditions.

Electrons in superconductors form Cooper Pairs below their critical temperature. These pairs of electrons are held together with phonouns, which are atomic-level collective excitations. Phonouns are similar to photons in that they also have particle-like properties [1].

Unconventional superconductors are still not well understood and remain an open question in Physics. Understanding them could lead to the discovery of superconducting materials stable at room temperature under atmospheric conditions. Such a material would have large implications, such as super efficient electricity transfer and vast efficiency improvements for applications like particle accelerators and power lines.

Machine learning can be an excellent asset in the search for such superconducting materials. An accurate model that can predict critical temperature based on composition attributes could be used to vet candidate materials before experimental testing, allowing experimentalists to focus on promising materials.

Calculating uncertainty is also important, as it can give research a confidence range for a machine learning model's prediction. This paper attempts to create an accurate model with uncertainty quantification, allowing experimentalists to predict critical temperatures with the confidence intervals.

II. METHODOLOGY

II.1. Database

We chose to use one of the most popular experimental datasets, the supercon database from Japan's National Institute for Materials Science. This datasets contains 16,414 superconductor chemical compositions and their experimentally measured critical temperatures. Unfortunately, the database is not currently available on their website for unspecified reasons, so we obtained the dataset from a github repository that used this data [2].

In this dataset, there are 10,154 samples with a critical temperature below 10K and 6,210 samples above 10K. There are 159 samples with temperatures above 100K and 0 samples above 260K. A histogram represents this information graphically in Figure 1. Since most of this data is at low temperatures, we can expect less accuracy in our models for higher critical temperature superconductors.

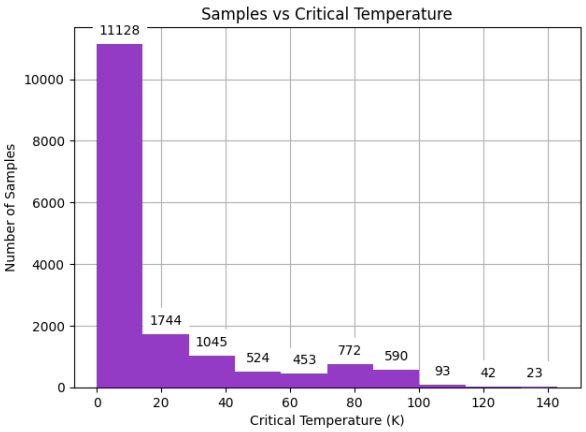

FIG. 1. A histogram of critical temperatures in the dataset. Most superconductors are found at low temperatures and this dataset is representative of that.

Most superconductor databases do not include enough information to train an effective machine learning model, but such data can be extracted from the data they do provide. We use matminer to produce our features from the provided material data. Matminer is a python library that generates data from various measured properties of a material [3]. Matminer collects existing calculations into a machine learning friendly python package.

Our database only provides the superconductor composition data. Matminer's featurizers can generate 53 features from the composition of a material. If we had band structure or other data, we could produce more information that we could use in our model.

II.2. Machine Learning

Previous papers have used random forest models to predict critical temperature [citation needed], but this paper will examine eight models before settling on two for further investigation. All models are implemented with Scikit-Learn, with the notable exception of a mlens

* Lycoming College, Williamsport, PA; klekirk@lycoming.edu

superlearner [4, 5]. We will also use MAPIE models for uncertainty, discussed in Section II.3.

We decided to test a variety of machine learning models to ensure we would find an optimal model. These models are listed below:

- Linear Regression
- Support Vector Regression (SVR)
  - SVR makes predictions with decision boundaries, which are lines parallel to the regression line or curve. The model aims to maximize the amount of data within the decision boundaries and has hyperparameters to modify sensitivity to prevent overfitting.[1]
- Elastic Net
  - Elastic Net uses L1 and L2 penalties to stabilize a linear model.
- Bayesian Ridge
  - Bayesian Ridge uses probability distributors instead of point estimates for a linear model.
- Decision Tree
  - Decision trees are very interpretable - they break predictions into nodes of the tree, eventually leading to a prediction value. These trees can be represented graphically and show how they produce results, unlike most machine learning models.
- KNeighbors (KNN)
  - KNN models are a little different, they store all the data and predict values based on a similarity measure. The model looks at a specified number of similar neighbors to produce a prediction.
- Random Forest Regression (RFR)
  - RFR is an ensemble method that uses numerous decision trees and subsamples the data with replacement. This means that the model replace data after using it in a subset.
- Extra Trees
  - Extra Trees is like RFR, but it does not replace the data after use in a subset.
- Superlearner
  - Our superlearner is a mlens model that can combine multiple high-scoring Scikit-Learn model predictions and sometimes improve the performance from the individual models.

[1] Overfitting occurs when a model is trained to be too specific to a particular dataset and is not generalizable.

II.3. Quantifying Uncertainty

Our evaluation functions can produce uncertainty calculations using forestci, mapie, or lolopy [6–8].

Forestci is python implementation of an algorithm from [9] that predicts confidence intervals for random forest models. It is the fastest of the uncertainty methods listed.

The Model Agnostic Prediction Interval Estimator (MAPIE) python library is more recent implementation of jackknife based on [10]. MAPIE uses various resampling methods. Most methods require the use of MAPIE's own MapieRegressor, which accepts an Scikit-Learn regressor and keeps track of uncertainty as the model is trained. We chose to use the MAPIE-plus method, which is based of the Jackknife+ algorithm. This is the default uncertainty method for MAPIE and is recommended by the developers.

MAPIE also has a prefit method, but it is difficult to extract uncertainty bars for individual points from this data - it splits a celebration set off the test set to generate uncertainty, so it can't be easily added to our plot of test set predictions. Thus, we will only compare the normal MAPIE methods with the other libraries. MAPIE trains much slower than other models, particularly on our superlearner, but it is still considerably faster than our final uncertainty model, lolopy.

Lolo is a Scala random forest machine learning library and is not a native python implementation. Lolopy is a python wrapper for lolo, but this implementation is very slow for large datasets. We were unable to successfully train a lolopy model due to these factors and time limitations of the project, but our code supports lolopy models.

II.4. Project Structure

The source code used for this paper is available publicly on github at `https://github.com/sylphrena0/classe2022`. This repository also includes the source files for this latex paper, data files, images, and documentation files.

Our research uses numpy and pandas throughout our code to handle arrays and tabular data [11, 12]. We also use matplotlib and seaborns to generate our graphs [13, 14].

The code is split into multiple python files so processes could be completed in stages and to maintain readability in the code. Most of our testing and final training was completed in juypter notebooks, but some computations were highly computationally expensive and needed to be run remotely. For these jobs, we created simple python files and made bash scripts to run them on Cornell's CLASSE compute farm. We also made several bash aliases and functions to simplify the compute farm workflow, which are also available on the github repository.

Since we used multiple files, we chose to create shared dependencies files where we defined functions to import

data, train models, and generate our graphs. These files are then imported in all the relevant scripts to reduce redundancy. More detailed explainations of the purpose of each file is available in the github readme file and documentation within the files.

II.5. Project Evaluation

First, the featurizer script imports the dataset, extracts features from the material compositions, and exports the csv data. This script is one of the most computationally expensive and takes several hours to run on the CLASSE compute farm with 64 dedicated cores.

After the features are exported, our analysis jupyter notebook imports the data with the shared import function and exports histograms and a correlation matrix.

Next, the training_single jupyter notebook or script can train individual models with the shared evaluation functions. This is used to get a landscape of initial performance before optimization. After training, the function plots the actual T_C versus the model prediction, using a heatmap to visualize the difference from the ideal prediction.

The optimizer script then uses a grid of manually defined hyperparameters to optimize models based on R2 score. This allowed significant improvements to baseline models. After optimization, the optimized models can be plotted in our single training notebook. After confirmation that the model is better than the baseline, the models can then be plotted together in a single graph using our bulk training notebook.

We evaluated our models using several metrics - R2 scores (R2) for regression evaluation, Mean Squared Error (MSE) and Mean Absolute Error (MAE) for error evaluation, and MAPIE Effective Mean Widths (MWS) for uncertainty evaluation.

III. RESULTS

III.1. Model Optimization

Each model's hyperparameters[2] were optimized with various optimization methods. Optimization can be computationally expensive, so we chose to optimize on a randomly selected subset of 2,000 materials, using a numpy random state for reproducibility. To start, we used Scikit-Learn's GridSearchCV, which tests combinations from a grid of hyperparameters and returns the best performing model based on a specified metric.

We also implemented Bayesian optimization using Gaussian Processes methods from the Scikit-Optimize library [15]. Bayesian optimization attempts to optimize models intelligently, instead of randomly testing specified hyperparameters. We provide a range of hyperparamters values for Bayesian optimization to test, and the algorithm uses acquisition functions to decide which specific values to use within the specified range. This is different from GridSearchCV, which is simpler but can take much longer to find optimal hyperparameters. We only implemented Bayesian optimization on our top models.

The performance of each optimizer on selected high performance models is shown in Table I. Note that the optimal method for each model (shown in bold), was found with a different method each time. The Bayesian optimization was much less computationally expensive than GridSearchCV, however, and is the recommended method for large datasets.

Model	Optimizer	R2	MSE	MAE	MWS
Random Forest	Base Model	0.807	140.482	5.791	53.135
	GridSearchCv	0.804	142.524	5.816	52.416
	Bayesian – PI	0.815	134.509	5.72	50.985
	Bayesian – EI	0.814	135.267	5.763	51.533
	Bayesian – gp_hedge	0.815	134.785	5.747	51.021
Extra Trees	Base Model	0.818	132.395	5.214	48.153
	GridSearchCv	0.818	132.374	5.215	48.231
	Bayesian – PI	0.818	132.104	5.225	48.154
	Bayesian – EI	0.816	133.5	5.257	48.757
	Bayesian – gp_hedge	0.819	131.783	5.202	47.732
KNN	Base Model	0.646	257.108	8.563	75.612
	GridSearchCv	0.703	216.186	7.515	69.758
	Bayesian – PI	0.652	253.137	8.201	74.177
	Bayesian – EI	0.652	253.137	8.201	74.177
	Bayesian – gp_hedge	0.566	315.238	8.159	84.546

TABLE I. Comparison of optimization methods by model.

III.2. Optimized Models

	Model	R2	MSE	MAE	MWS
Unoptimized	**Extra Trees Regression**	0.818	132.395	5.214	48.153
	Random Forest Regression	0.807	140.482	5.791	53.135
	KNeighbors Regression	0.646	257.108	8.563	75.612
	Decision Tree Regression	0.644	258.945	6.886	76.715
	Bayesian Regression	0.392	441.925	14.52	90.7
	Linear Regression	0.392	442.163	14.506	90.705
	Elastic Net Regression	0.328	488.242	15.603	97.938
	Support Vector Machines	0.084	666.16	15.511	134.628
Optimized	**Extra Trees Regression**	0.819	131.624	5.205	48.088
	Random Forest Regression	0.816	133.87	5.714	50.759
	KNeighbors Regression	0.703	216.186	7.515	69.758
	Decision Tree Regression	0.664	244.095	7.152	74.834
	Elastic Net Regression	0.392	442.133	14.487	90.713
	Bayesian Regression	0.392	441.925	14.52	90.7
	Linear Regression	0.392	442.163	14.506	90.705
	Support Vector Machines	0.325	490.661	14.186	106.946

TABLE II. CV scores of the models using all features and data, comparing optimized and base hyperparameters.

Our base models provided good results and the top two models, Random Forest and Extra Trees, only improved their scores marginally with optimization. Our numerical

[2] Hyperparameters are machine learning parameters that change how a model is trained.

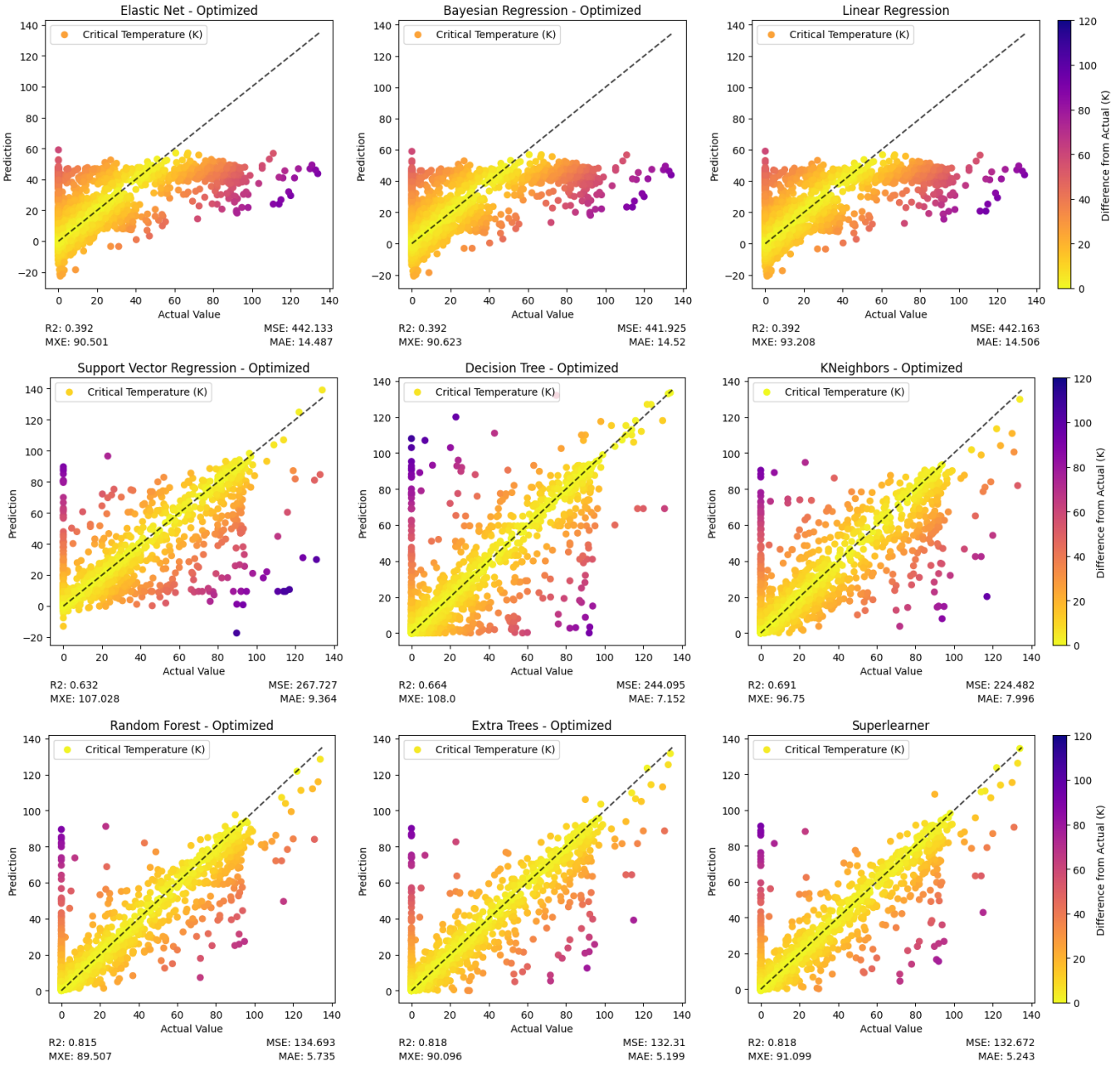

FIG. 2. Graph of optimized model predictions versus actual critical temperatures for each base model. The dotted line represents the optimal prediction for any given critical temperature.

results of the optimized and optimized models are shown in Table II.

Our linear models performed poorly, with none of these models exceeding an R2 value of 0.4, and with MAE above 14K. Additionally, uncertainty is quite high for these models, with the confidence interval widths (MWS) exceeding 90K. The decision tree and KNeighbors models perform a little better, with R2 values around 0.65 and MAE around 7K. Our ensemble models stole the show—the extra trees model consistently performed the best, with an R2 value of 0.819, MAE around 5K, and confidence intervals around 50K. Random Forest had marginally worse R2 and MAE scores. These results are shown graphically in Figure 2.

As shown in Figure 2, the linear models do not make good predictions at high critical temperatures. All the linear models make very similar predictions - the ensemble methods and KNeighbors have better predictions and are less uniform.

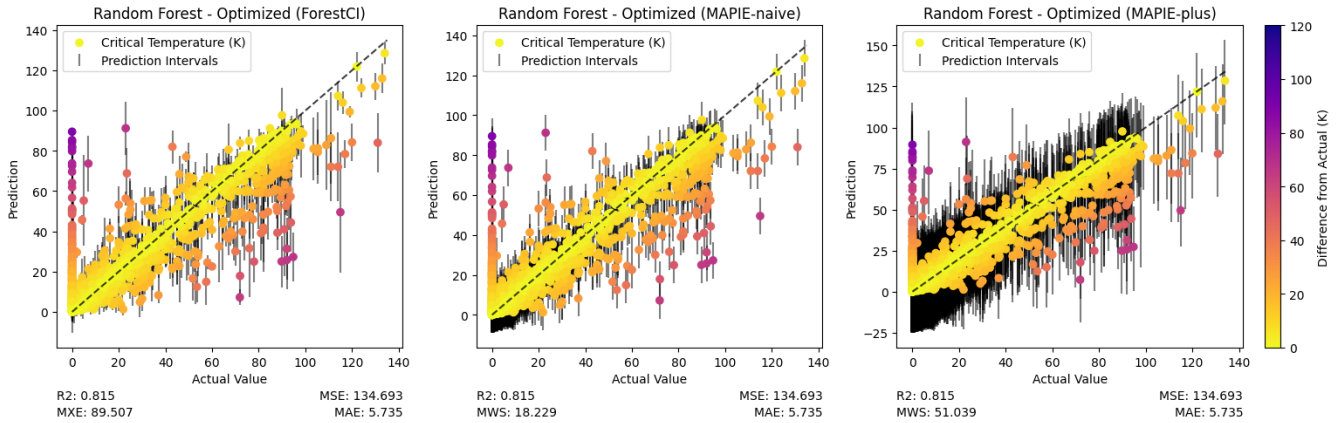

FIG. 3. Comparision MAPIE methods and ForestCI uncertainty calculations for Random Forest models with plot of optimized model predictions versus actual critical temperatures. Note that all other uncertainty calculations in this paper use the MAPIE-plus method, which uses a Jackknife+ algorithm. This is the default method for MAPIE.

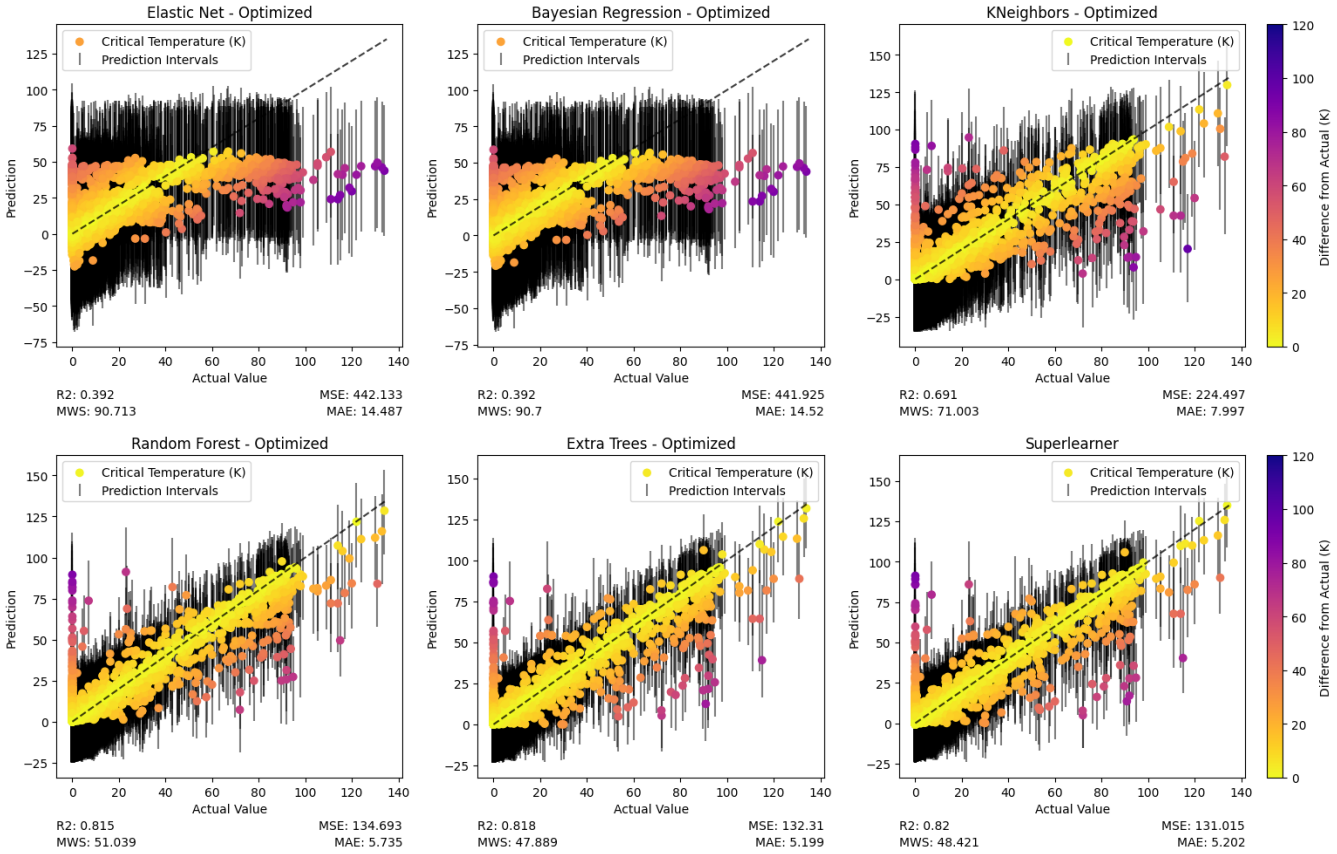

FIG. 4. Graph of optimized model predictions versus actual critical temperatures for selected base model with confidence intervals. Other models have similar invervals but are omitted here for brevity.

III.3. Comparing Uncertainty

As discussed in Section II.3, we quantified uncertainty for our predictions, mainly using MAPIE. Since our other uncertainty libraries only work for Random Forest Models, we used MAPIE to evaluate our models. We trained a Random Forest Model model with ForestCI to compare with MAPIE methods of uncertainty quantification.

As shown in Figure 3, MAPIE's naive confidence interval method has very similar to the ForestCI calculations. MAPIE has other prediction methods (base and minmax) that produce somewhat similar results to the MAPIE-plus (Jackknife+) method, but since we are not using those algorithms, we chose to compare our selected MAPIE method and the ForestCI-like method to compare to ForestCI.

ForestCI and MAPIE-naive produce very similar results, with MAPIE-naive making slightly more optimistic prediction intervals. However, MAPIE-plus produces prediction intervals which intersect with the actual value much more reliably. Thus, MAPIE-plus may be a better choice for our models, regardless of ForestCI's limited scope (only Random Forest Regression models).

III.4. Uncertainty Models

Since MAPIE is model agnostic, we have uncertainty predictions for all our models. We discussed our chosen MAPIE numerical uncertainty metric in Section III.2, but we can also represent the prediction interval graphically, showing the prediction interval for each predicted point. This is shown in Figure 4.

In Figure 4, each point has error bars that show the confidence interval. The model can be visually evaluated by seeing how many of the error bars intersect the dotted line, which represents the ideal prediction. While smaller prediction intervals are great, a model is not useful if it most predictions are not within error of the actual value. The bottom three models, ensemble models, have the smallest error bars and most prediction points have the actual value within the prediction interval.

The linear models in Figure 4 have unacceptablely large prediction intervals, in addition to their poor R2 and error metrics. KNeighbors is slightly better, but the uncertainty is still very high. Once again, our ensemble methods show the best metrics, with much smaller prediction intervals that still almost always include the actual value.

III.5. Cleaned Models

Introduction/Description of how we cleaned data

III.6. Feature Selection

Our

IV. CONCLUSION

IV.1. Future Work

ACKNOWLEDGMENTS

I would like to thank my mentor, Suchismita Sarker, for her guidance and support with this project. This work is supported by the U.S. National Science Foundation under award number NSF PHY-2150125, REU Site: Accelerator Physics and Synchrotron Radiation Science.

[1] J. W. Rohlf, Superconductivity, in *Modern Physics: From Alpha to Z* (Wiley, 1994) Chap. 15.
[2] vstanev1, Vstanev1/supercon: Data used in "machine learning modeling of superconducting critical temperature" paper (2018).
[3] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, Computational Materials Science **152**, 60 (2018).
[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).
[5] S. Flennerhag, Ml-ensemble (2017).
[6] K. Polimis, A. Rokem, and B. Hazelton, Journal of Open Source Software **2** (2017).
[7] V. Taquet, G. Martinon, N. Brunel, I. Ibnouhsein, F. Deheeger, R. Adon, A. Papp, A. A. Goumbala, A. Borgohain, T. Morzadec, and et al., Mapie - model agnostic prediction interval estimator (2022).
[8] M. Hutchinson, Citrineinformatics/lolo: A random forest library (2022).
[9] S. Wager, T. Hastie, and B. Efron, Journal of machine learning research : JMLR **15**, 1625 (2014).
[10] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, Predictive inference with the jackknife+ (2019).
[11] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, Nature **585**, 357 (2020).
[12] T. pandas development team, pandas-dev/pandas: Pandas (2020).
[13] J. D. Hunter, Computing in Science & Engineering **9**, 90 (2007).
[14] M. L. Waskom, Journal of Open Source Software **6**, 3021 (2021).
[15] T. Head, M. Kumar, H. Nahrstaedt, G. Louppe, and I. Shcherbatyi, scikit-optimize/scikit-optimize (2021).