Santiago Garcia, Christopher

June 8, 2025

Final Project Report

Introduction/Motivation

Superconductors are materials that conduct electricity with zero resistance. These materials exhibit the Meissner Effect which is when a superconductor expels the magnetic field and creates its own current [1]. This expulsion allows for a magnet to levitate above a superconductor, seen in Figure 1.

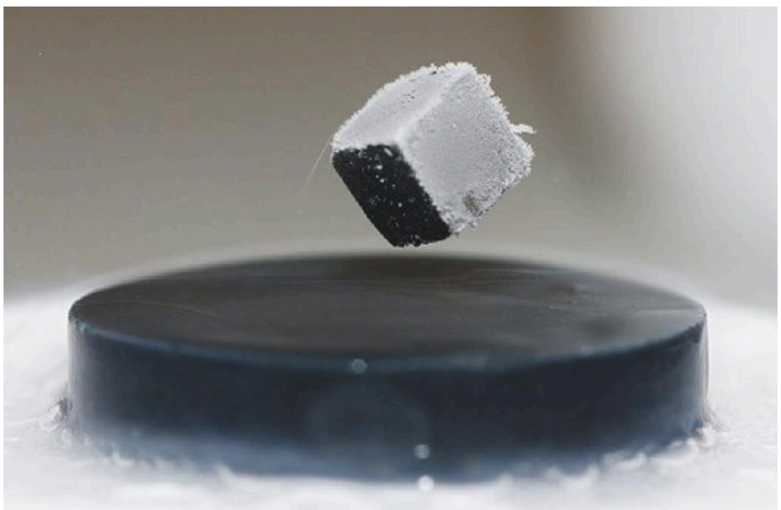

Figure 1. An equal and opposite field caused by Meissner Effect [2]

The discovery of this effect made it clear that superconductors were a unique phenomenon. The superconductor phenomenon was thought to be completely explained by the BCS theory, which states that superconductivity is caused by electron-photon interaction [1]. Eventually, limitations within the BCS theory were found once high critical temperature (T_C) superconductors were discovered. High T_C materials are superconducting at a higher temperature, above 77 K [2]. This boundary is important because 77 K is known as the temperature at which liquid nitrogen (LN2) can exist at. The importance of liquid nitrogen will be discussed in the next section of this paper. Superconductors are essential because they are used in magnetic resonance imaging (MRI), quantum computers, and power grids [3]. This paper will mainly focus on the importance of high critical temperature superconductors in MRI machines.

In 2018, Kam Hamidieh used XGBoost to predict the T_C in the article "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor" [4]. The analysis discussed in this paper differs from Hamidieh because this work uses different linear models for interpretability. This work also uses support vector machines to include classification towards the dataset.

Problem Definition

Superconductors are relevant to the medical field because they allow doctors to make diagnoses from the detailed images MRI machines provide. For MRIs, superconductors are twisted in a coil (Figure 2) and cooled with liquid helium to reach their critical temperature and generate a strong magnetic field which is aligned with protons in the body to achieve images [3,5].

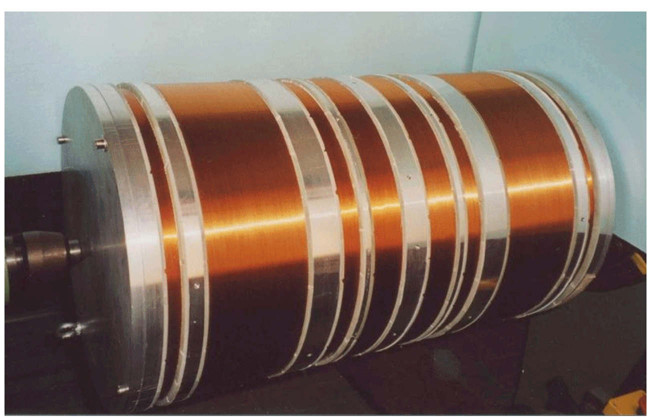

Figure 2. Image of superconducting material in a coil [6]

There are various issues with using liquid helium to cool superconducting materials such as high cost, scarcity, and high maintenance. Helium is a scarce element because there is a limited number of natural sources which causes the price of helium to skyrocket from $7.57 per cubic meter to $14 in the last couple of years [7, 8]. As the demand for helium increases, so will its price. Helium is difficult to maintain because it flows through microscopic holes and eventually leaks into the atmosphere [7]. These three issues with liquid helium are all interrelated. To reduce costs and improve usability, hospitals should switch to MRI machines that utilize high critical temperature superconductors cooled with liquid nitrogen. High critical temperature materials are able to exhibit superconductivity at higher temperatures, decreasing the depancy on helium. Liquid nitrogen is significantly cheaper and readily available compared to its coolant competitor. This paper is not pushing for LN2 based MRI machines because low critical temperature materials would not reach superconductivity. The specific data analytic problem addressed in this paper will use machine learning models to: (1) predict the critical temperature of superconductors and (2) classify these materials into high T_C versus low T_C.

Methods

The data used to solve problems (1) and (2) comes from Hamidieh's work in "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor". This paper included two files, but only train.csv was used for this paper's work because it included 21,263 superconductors each with 81 elemental properties. Hamidieh's analysis solves problem (1) using XGBoost, but this paper will primarily use Ridge and Lasso regression models. XGBoost is re-implemented into this work to provide insight into which model is best suited for this task. This paper solves problem (2) using Support Vector Machines (SVMs).

Ridge regression produces a model that includes all features and adds a penalty function to the loss function [8]. On the other hand, Lasso regression produces a model that only keeps the most significant features [8]. These two regression models were selected because they reduce overfitting and are able to handle many correlated features. XGBoost uses decision trees to perform gradient boosting, improving weak models by combining it with other models [9].

Support vector machines were chosen to address problem (2) because they are well suited for binary classification, low T_C versus high T_C, and are effective in high dimensional spaces [10]. Support vectors, points closest to the hyperplane, can be visualized in Figure 3.

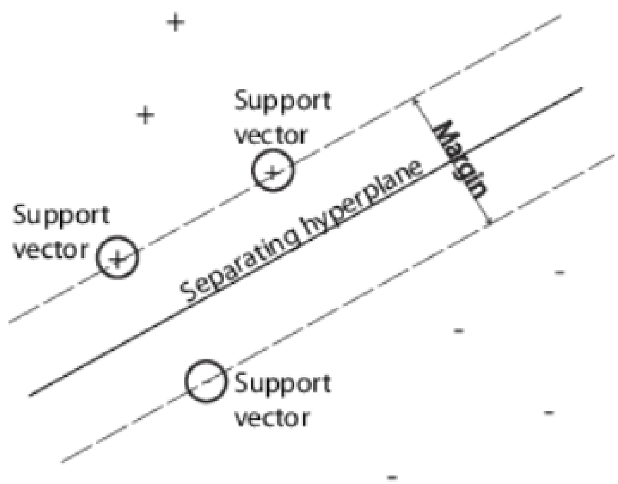

Figure 3. The hyperplane separates the two classes and is chosen based on the largest margin [10].

This classification decides whether or not the material can be cooled using liquid nitrogen.

Python programming language was used for this project. AI tools, specifically ChatGPT, was used to understand certain functions and to understand why the code was not compiling at times. The software packages used were pandas, numpy, matplotlib, scikit-learn, and xgboost. The pandas and numpy packages allow data to be analyzed and manipulated. For data visualization, the matplotlib package was used. The scikit-learn package contains tools for classification, regression, clustering, and dimensionality reduction. To reimplement Hamidieh's analysis into this work, the xgboost package was added. These packages were called in with the import or from function. The workflow behind the software package can be found as comments above the lines of code. This was done instead of explaining in this report to avoid redundancy and confusion. Moreover, to speed up the analysis, the program was separated into two Jupyter cells. Prior to the separation, the code would compile for too long and made debugging difficult. Charts that include the predicted critical temperatures for each superconductor and the actual critical temperature are submitted to BruinLearn.

Results

To quantify the accuracy of each models' predictions of T_C, root mean squared error (RMSE) and the coefficient of determination (R^2) were calculated, and placed in Table 1.

Table 1. RMSE and R^2 values for each predictive model.

Model	**RMSE**	$\mathbf{R^2}$
Ridge	17.61	0.73
Lasso	18.46	0.71
XGBoost	10.02	0.91

RMSE uses Euclidian distance to show how far predicted values are from the actual T_C [11]. An RMSE value of 0 is best, but lower values mean the model was able to accurately predict T_C. XGBoost had the lowest RMSE value, 10.02, and was closer to the actual T_C by a value of at least 7.59. R^2 shows the proportion of variance and ranges from 0 to 1. A R^2 value closer to 1 means the predictions are closer to the true value. Similar to the RMSE, XGBoost had the highest R^2, 0.91, and was at least 0.18 away from the next nearest model.

Figure 4 is a scatter plot showing the predicted versus actual T_C for Ridge and Lasso Regression. The data points stray away from the ideal y = x line.

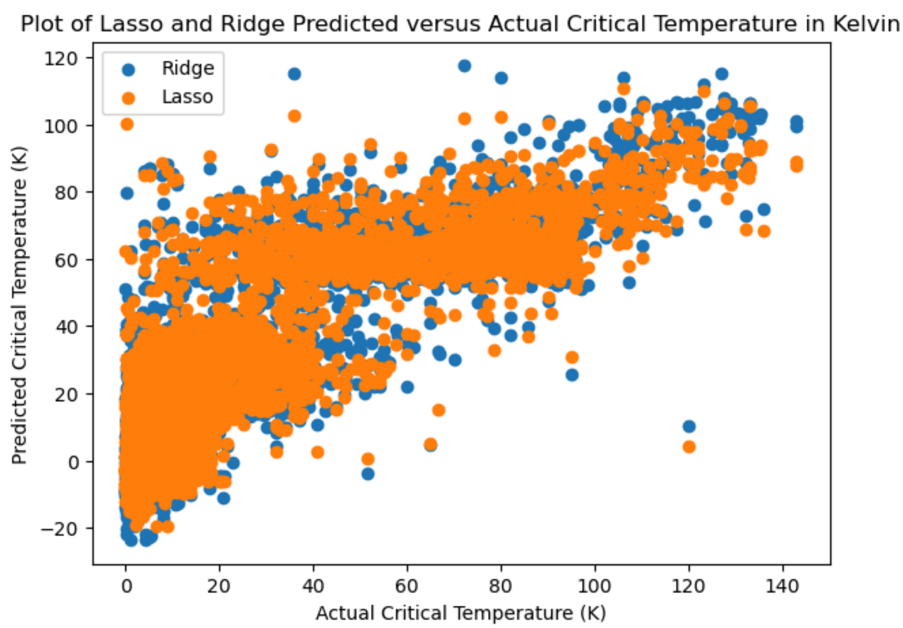

Figure 4. Predicted vs actual T_C for Lasso and Ridge regression.

Figure 5 is a scatter plot showing the predicted versus actual T_C for XGBoost. The data points lie tightly around the ideal y = x line and support the RMSE and R^2 values in Table 1.

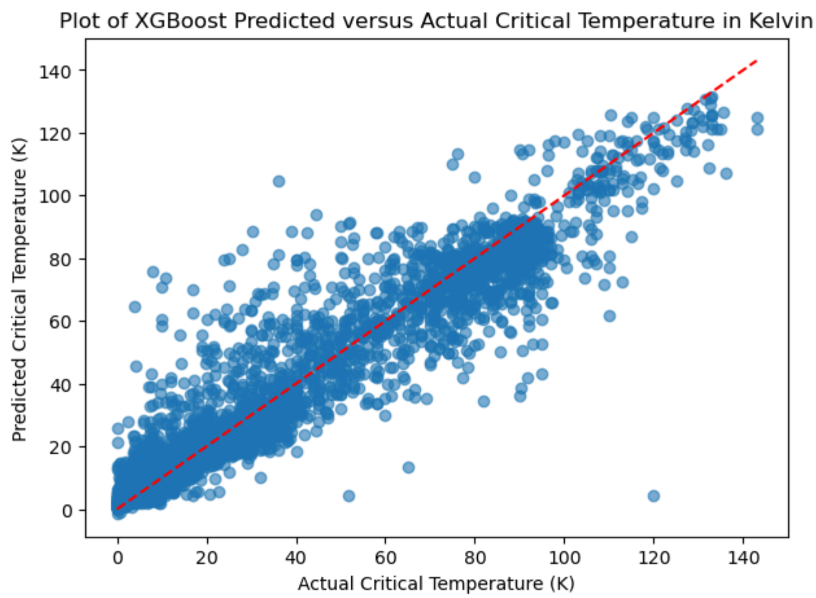

Figure 5. Predicted vs actual T_C for XGBoost.

The results from the Support Vector Machine classification are shown in Figure 6. From the precision score, the SVM reported a precision of 0.69. This means that 69% of the predicted high T_C were in fact high T_C superconductors.

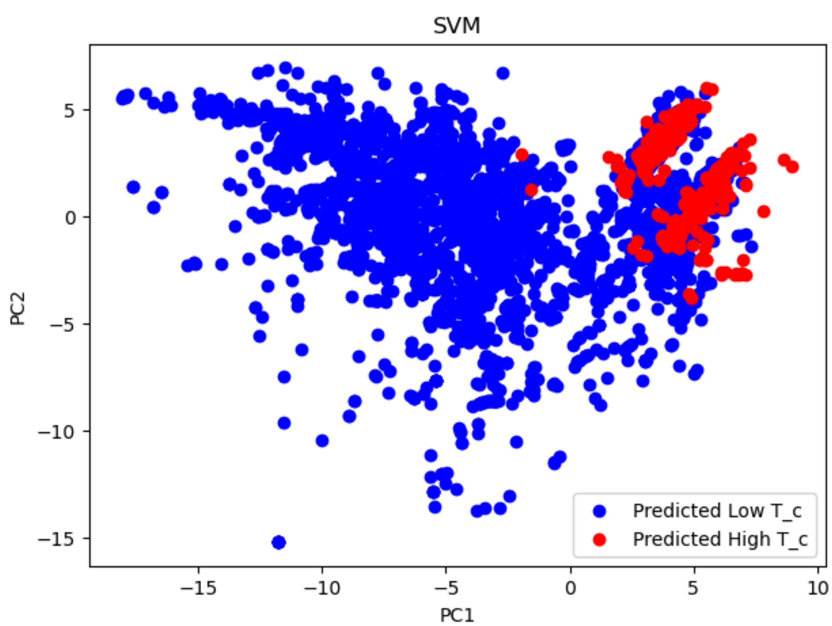

Figure 6. SVM classification results

XGBoost outperformed both Ridge and Lasso regression, indicating that Hamidieh's original model selection was the most optimal. The Support Vector Machine had a precision of 69%, which shows that the model was more likely than not going to get the prediction right. This is not good enough for MRI machines as they are expensive and the equipment cannot undergo trial uses.

References

[1] Solymar, L., Walsh, D. and Syms,R.R.A. (2014) Electrical Properties of Materials. Oxford University Press, Oxford, 484.

[2] Department of Energy. (n.d.). DOE Explains...Superconductivity. Retrieved from Energy.gov website: https://www.energy.gov/science/doe-explainssuperconductivity

[3] “Drawn to Superconducting Magnets | Research UC Berkeley.” *Vcresearch.berkeley.edu*, vcresearch.berkeley.edu/heising-simons/jame-analytis/drawn-superconducting-magnets.

[4] Hamidieh, K. (2018). Superconductivty Data [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C53P47.

[5] Aarnink, René, and Johan Overweg. “Magnetic Resonance Imaging, a Success Story for Superconductivity.” *Europhysics News*, vol. 43, no. 4, July 2012, pp. 26–29, www.europhysicsnews.org/articles/epn/pdf/2012/04/epn2012434p26.pdf, https://doi.org/10.1051/epn/2012404.

[6] Imaging, P. (2022, February 16). How to Prepare for an MRI | Premier Imaging. Retrieved from Premier Imaging | Diagnostic Radiology website: https://premierimaging.org/how-to-prepare-for-an-mri/

[7] Bains, Callum. “The World Keeps Running out of Helium. There Is Now a Race to Prepare for the next Shortage.” *Bbc.com*, BBC, Apr. 2025, www.bbc.com/future/article/20250331-why-helium-shortages-are-worrying-the-world.

[8] “Helium Statistics and Information | U.S. Geological Survey.” *Usgs.gov*, 2012, www.usgs.gov/centers/national-minerals-information-center/helium-statistics-and-information.

[9] Nvidia. “What Is XGBoost?” NVIDIA Data Science Glossary, 2024, www.nvidia.com/en-us/glossary/xgboost/.

[10] Support Vector Machines for Binary Classification - MATLAB & Simulink. (n.d.). Retrieved from www.mathworks.com website: https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

[11] C3.ai. "Root Mean Square Error (RMSE)." *C3 AI*, c3.ai/glossary/data-science/root-mean-square-error-rmse/.