Domain Background:

This project is an investigation into superconductivity. Certain materials known as superconductors when they are cooled below a characteristic critical temperature expel certain magnetic flux fields and have exactly zero electrical resistance. The critical temperature is the temperature at which the material loses all its electric resistance. Superconductors have many applications with one of the most popular and widely used being their application in Magnetic Resonance Imaging (MRI) which is widely used by health care professionals in creating and analysing internal body image's.

Typically, a material reaches superconductivity status when cooled to an extremely low temperature. Reaching these temperatures can be costly to implement and ultimately maintain to optimally use the material as a superconductor. As such, there is a great need to be able to identify where exactly a material will reach its critical temperature. It is the intention of this project to construct a model for predicting based on a number of input features the critical temperature of a material. This paper does not comment on the effectiveness of the material as a superconductor, more so it is looking to reduce the scope of potential materials for researchers and practitioners when looking for low-temperature superconductors. Additionally, it is hoped that the results discovered here will assist in the combination of materials to act as superconductors at higher temperatures.

Previous work in this area has been performed by papers published by *(Hamidieh, 2018)* and *(Curtarolo, et al, 2017).* They found success in applying machine learning algorithms in predicting the critical temperature of materials. The data-set is from the paper published, *(Hamidieh, 2018),* in which the features of a material were established to try and determine the critical temperature of a material given a number of features such as atomic mass, density and thermal conductivity.

Problem statement:

The problem here is, from a number of input features of a material, can the critical temperature of it be accurately and consistently predicted. In essence, it is a straightforward regression problem with a number of feature variables and a single target variable to be predicted, critical temperature. Work performed by *(Hamidieh, 2018)* found success using an XGBoost model to predict the critical temperature ultimately resulting in a mean square error of 9.5K. These results will be compared to ones in this project to gauge the success and practicality of the constructed model.

As noted, this is a regression problem in which supervised learning techniques will be applied. It is proposed to use a Random Forest regression to try and surpass the work of *(Hamidieh, 2018).* Additionally, to differentiate the work, a neural network will be constructed to see if consistent prediction results can be achieved via such a model. Previous work in this area focused on supervised learning techniques, is it hoped that utilising deep learning superior results can be achieved.

Description of the data-set:

The data-set consists of eighty-one features with the intention of predicting the critical temperature of superconductors. As discussed, the data-set comes from a paper by *(Hamidieh, 2018).* This is a relatively straight-forward data-set in terms of complexity, in that there are a

number of feature variables with the intention being to predict a single target variable. As discussed in *(Hamidieh, 2018)* the data-set was obtained from NIMS, a public institution in Japan. The final data-set used in *(Hamidieh, 2018)* consists of 21,763 rows of materials and their critical temperatures. This is appropriate for such an investigation as we are a constructing a model to predict critical temperature based on given input features. Additionally, through the use of an identical data-set we can accurately compare our final results and ultimately be able to show if the constructed model returns superior results or not.

Define what the proposed solution for the problem outlined above:

(Hamidieh, 2018) mentions how random forests were not used due to the length of time to train given the data-set. It is proposed here to utilise Lasso regression techniques, *(Fonti, 2017)* to aid in reducing the number of features prior to training and testing constructed models. *(Curtarolo, et al, 2017)* utilised random forests, although on a much smaller data-set. It is hoped that through employing feature selection techniques training and testing can be quicker and ultimately make for a more useful and applicable model.

As mentioned above, it is proposed to create a neural network with a final linear activation function layer and a square error loss cost function to measure the final predicted values in addition to the random forest. Through calculation of this error function we can compare our results both to the Random Forest constructed in this project and the XGBoost model proposed in *(Hamidieh, 2018).*

If there is any historical models that produced results which can be leveraged for comparisons:

It is proposed to use the results of *(Hamidieh, 2018)* to compare and contrast the final results of our constructed model. Ultimately, they found they finished with a R-Squared score of 0.92 and a root mean square error of 9.5K indicating that in terms of predicting power there model tended to be off by about 9.5K when predicting critical temperature when using the XGBoost model. When the multiple regression model was used in *(Hamidieh, 2018)* the scores were 0.74 and 17.6K for R-squared and root mean square error respectively. It can be seen that the XGBoost offered a much greater prediction accuracy than the multiple regression model .The aim will be to improve on these scores

In order to evaluate the model, the root mean square error will be used shown in the below formula. This will allow us to compare our final results to those seen in *(Hamidieh, 2018)* to see if the model offers greater prediction accuracy.

Average of (observed-predicted) ^2

Outline of Project design:

As discussed above, it is proposed to use the data-set from work performed by *(Hamidieh, 2018).* As we ultimately want to be able to compare and contrast models in addition to the data-set having already been pre-processed to such an extent that successful models have been built from it, no pre-processing on the data will be performed. The data-set will be taken as is into the processing stage where it will be normalised and investigated at a base level to see if any inter-dependencies or relationships exist between the features which can serve to aid us going forward at the training and testing stage.

As mentioned, this data-set has 81 feature variables to predict a single target variable. Previous work, noted that certain models were avoided due to the time taken to train the model due to the size of the data-set. With such a large amount of features it is anticipated that a number of them may act as noise for which inclusion may hinder the model rather than allowing it to perform to its full potential. With that in mind, it is proposed here to invest a significant amount of time and effort into understanding the features looking to employ Lasso regression techniques to aid us in selecting which combination of features will be used to construct the final model. Through reducing the dimensionality of the feature-set, it is hoped that will allow for quicker training and testing whilst also avoiding overfitting the training data-set.

To complement the work performed by the Lasso regression using a tree-based estimator will be implemented to help us with feature importance and information gain, tree based feature selection will be employed. This will allow us to truly capture entropy and information gain as we navigate through the tree. Ultimately, we want to obtain an understanding of which features contribute the most to predicting the critical temperature of a material. Following this, we will look to training the model. Training will be performed extensively here, with a number of parameters to tune and augment in order to find the correct balance between predicting and overfitting.

In addition to splitting the overall data into training and test, the training data itself will be split into training and validation sets. The ultimate aim here is to reduce variance both in training and in testing and it hoped that through the use of a validation set that it may assist in this. Additionally, by continuously validating our training results we can gain greater insight into tuning the model and hopefully strike a balance in our model i.e. over vs under fitting our data-set. There are 21,263 rows of data which will be split into an 80-20 training and test split. *(Hamidieh, 2018)* utilised a 2/3 vs 1/3 train vs test split. It is proposed here to use an 80-20 split both to differentiate out work and also to see if a larger training set will offer greater prediction ability and a superior model for determining the critical temperature.

As proposed, Random Forest's and Neural Networks will be implemented here to see if through their use can superior performance be achieved when compared to the XGBoost model proposed by *(Hamidieh, 2018).* The inclusion of random forests is to offer a direct comparison to previous work performed in the area as these type of decision tree based models have shown success. It is hoped that dimensionality will be reduced through using a limited number of features as proposed in the feature selection steps above.

The creation of a Neural Network, is to offer a different approach than what has been previously proposed. It is hoped that through the creation of a smart and succinct architecture that dimensionality will be reduced in the hidden layers and will offer a superior model to predict critical temperature than decision tree based model previously used. Instead of a Relu activation function, a linear one will be used.

As discussed, the final results following testing will be the calculation of the R-squared score and the root mean square error. This will allow us to compare our results to *(Hamidieh, 2018)* and be able to determine the viability of Random Forests and/or Neural Networks to predicting the critical temperature of materials.

References:

Hamidieh, Kam, A data-driven statistical model for predicting the critical temperature of a superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346-354

Valentin, S., Oses, C., Kusne, A. G., Rodriguez, E., Paglione, J., Curtarolo, S., and Takeuchi, I. (2017). Machine learning modeling of superconducting critical temperature. https://arxiv.org/abs/1709.02727.

Fonti, Valeria, and Eduard Belitser. "Feature selection using lasso." VU Amsterdam Research Paper in Business Analytics(2017).