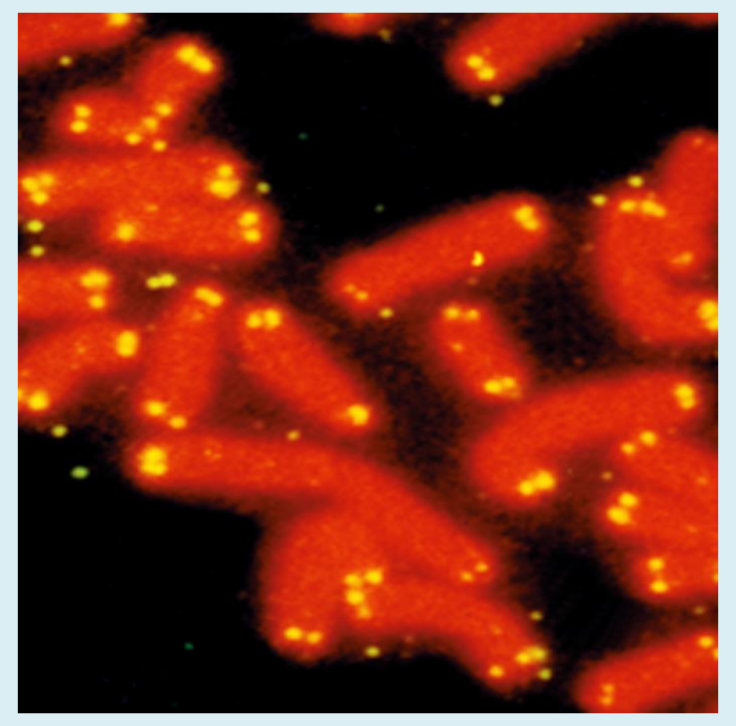
TELOMERES IN THE TELSEQ AND TELOMERECAT PIPELINES WITH MACHINE LEARNING ANALYSIS FOR BETTER PREDICTIONS

MICHALIS MYLONAS

SCHOOL OF MEDICINE CARDIFF HEATH CAMPUS CF14 4ER

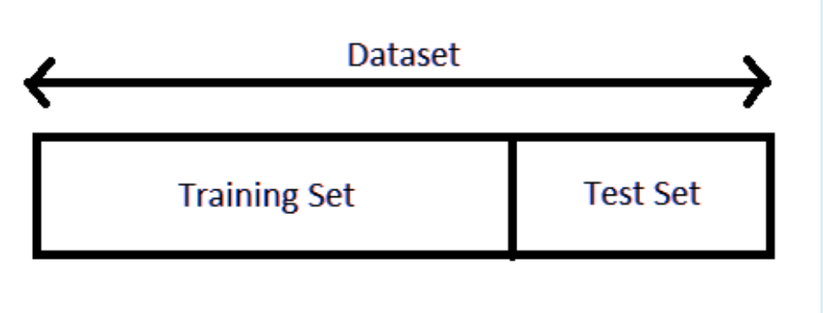
ABSTRACT

Telomeres are located at the end of each chromosome and they protect them from sticking to each other. Telomere length is important as it can help predict cancer and it can be related to ageing. This study has explored novel pipelines such as TelomereCat and TelSeq to improve telomere length predictions. This resulted in a much better prediction accuracy that eventually will be very useful in improving breast cancer predictions.



INTRODUCTION

- Located at the end of each chromosome
- Protect our genetic data
- Hold secrets on how we age and get cancer
- Each time they replicate they get shorter
- More Accurate assessment of the learning Algorithm.
- Used all data for training and for testing.
- STELA is a robust method for measuring telomere length



METHODS

- Scikit-Learn
- Supervised learning algorithms
  - Linear regression
- Cross-Validation
- Outlier Detection

IMPLEMENTATION

- Python
- PyCharm
- Anaconda
- TelSeq
- TelomereCat

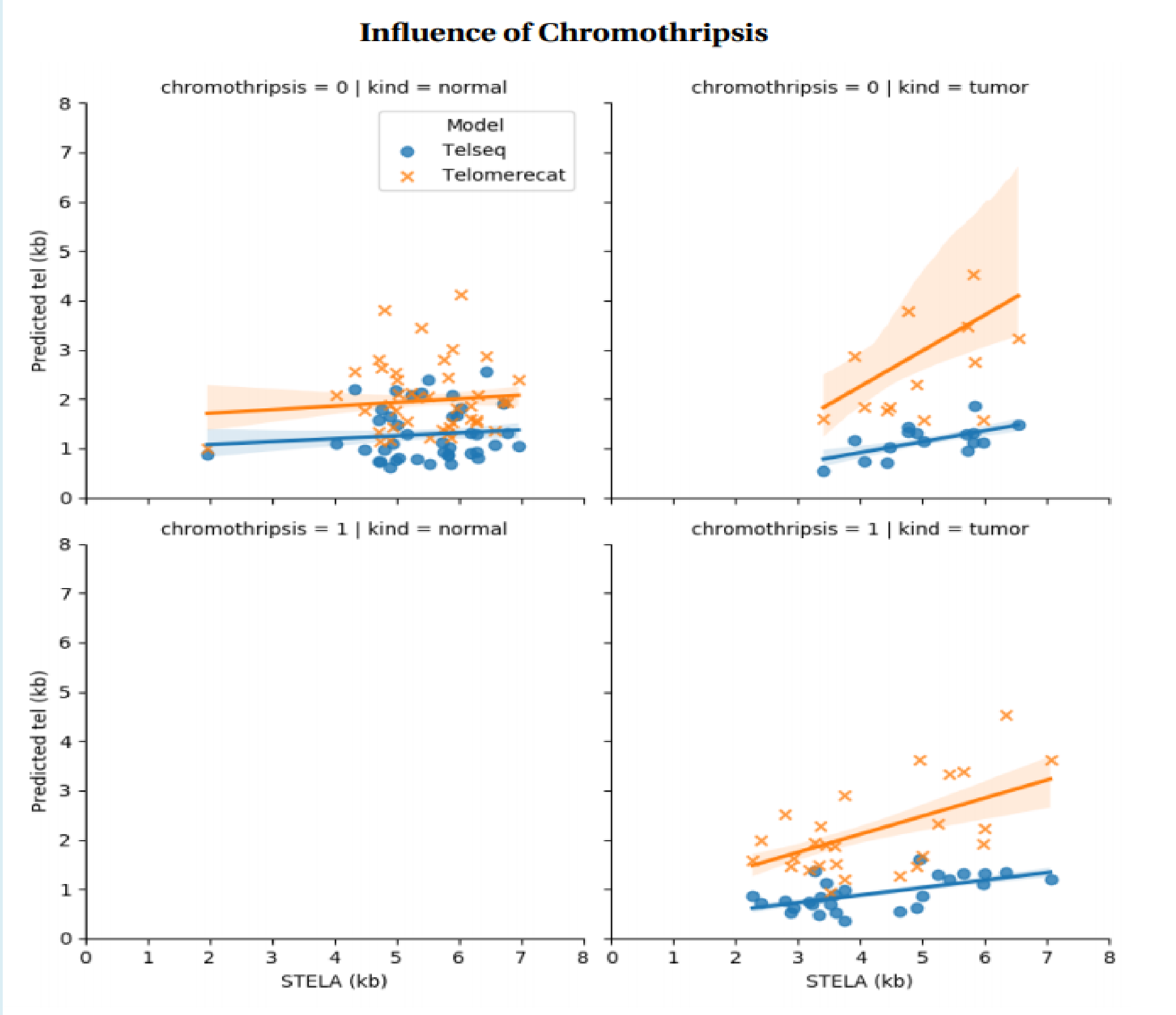
RESULTS

OUTPUT

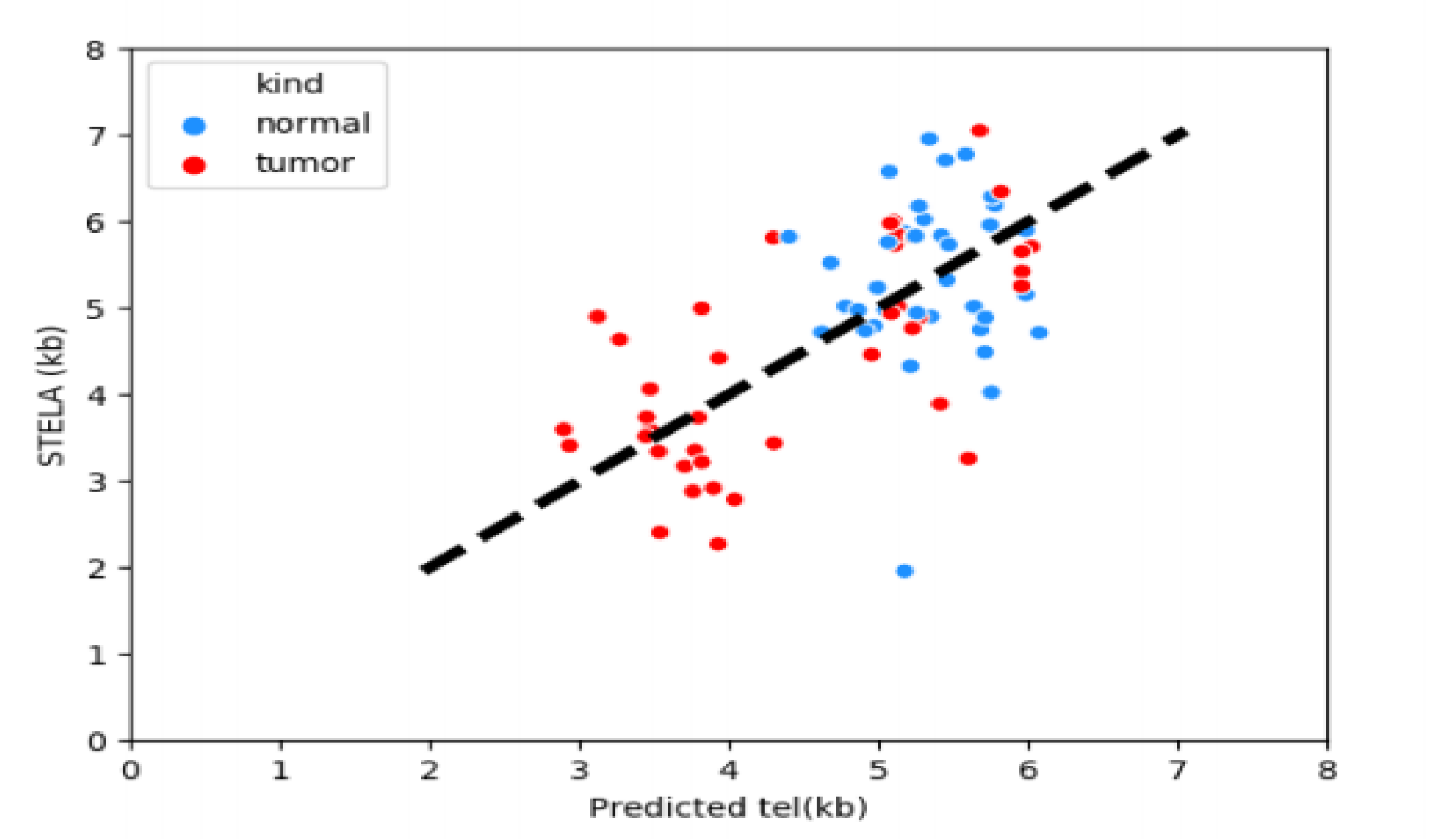
- Regression and cross validation improved prediction accuracy
- New parameters were useful
  - Chromothripsis
  - Binary tumour variable

Description	Initial	Final
R Square	0.1750	0.4186
RMSE	1.0663	0.8863

CHROMOTHRIPSIS

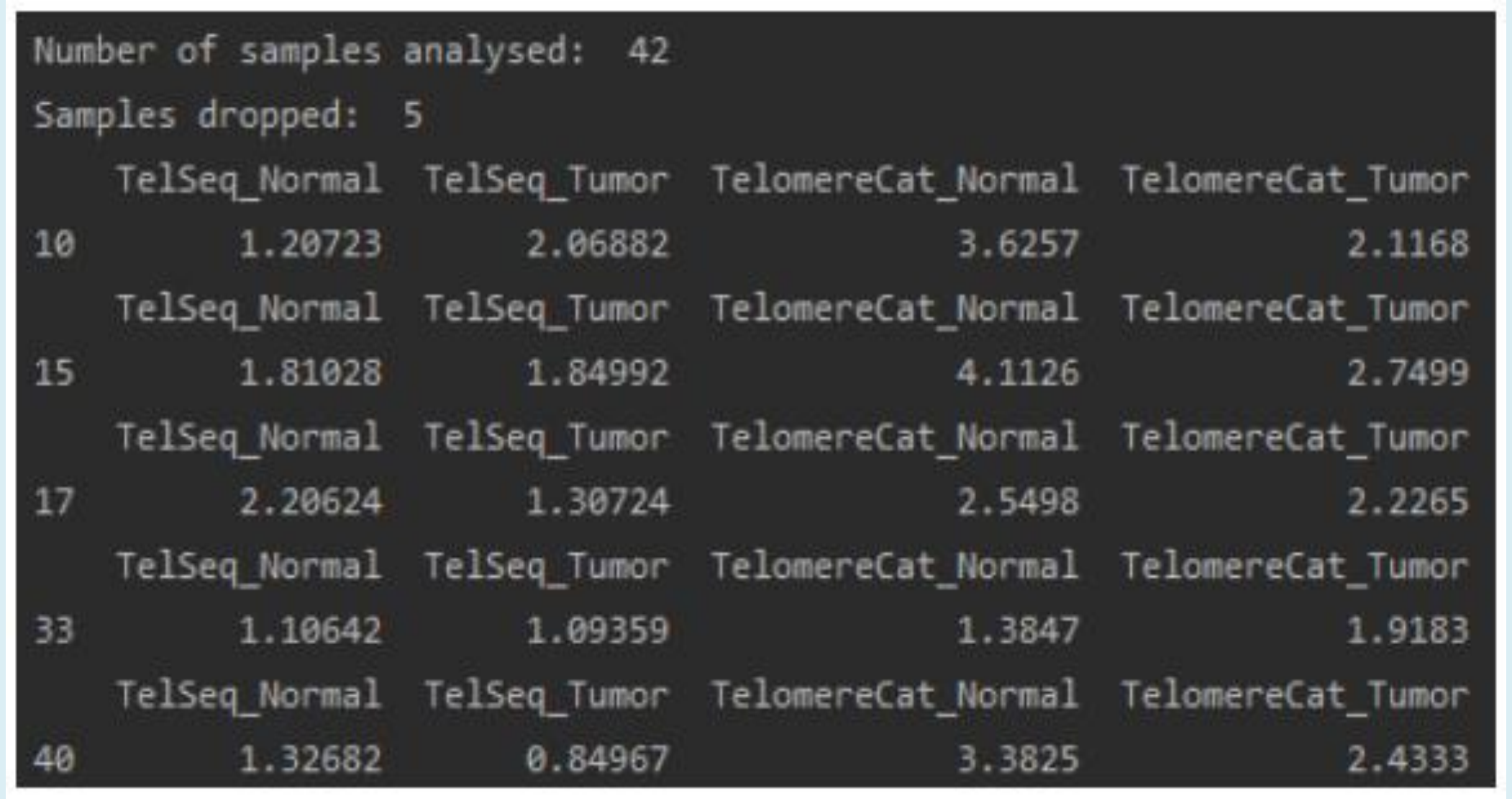


K-FOLD



OUTLIER DETECTION

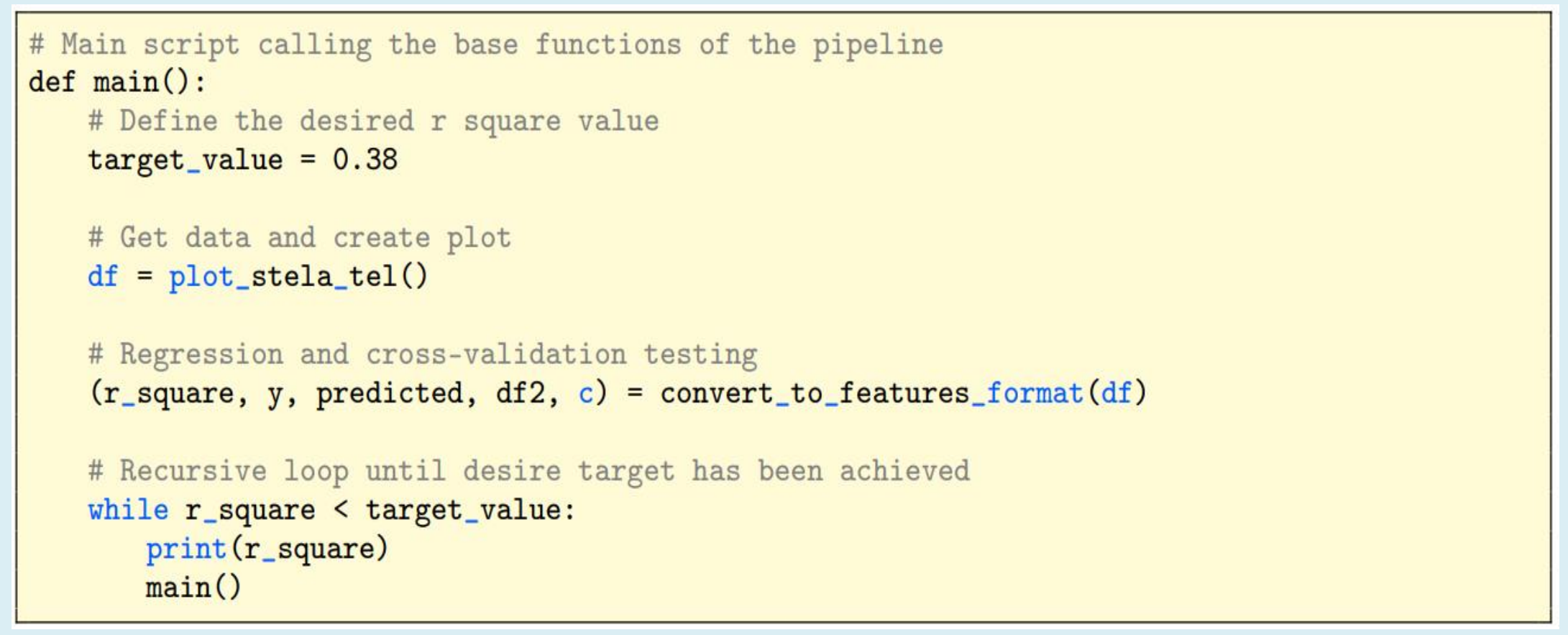
- Identification of anomalies
- Identify all the extremes
- Get rid of them or evaluate them



```
Number of samples analysed:  42
Samples dropped:  5
    TelSeq_Normal  TelSeq_Tumor  TelomereCat_Normal  TelomereCat_Tumor
10        1.20723       2.06882              3.6257             2.1168
    TelSeq_Normal  TelSeq_Tumor  TelomereCat_Normal  TelomereCat_Tumor
15        1.81028       1.84992              4.1126             2.7499
    TelSeq_Normal  TelSeq_Tumor  TelomereCat_Normal  TelomereCat_Tumor
17        2.20624       1.30724              2.5498             2.2265
    TelSeq_Normal  TelSeq_Tumor  TelomereCat_Normal  TelomereCat_Tumor
33        1.10642       1.09359              1.3847             1.9183
    TelSeq_Normal  TelSeq_Tumor  TelomereCat_Normal  TelomereCat_Tumor
40        1.32682       0.84967              3.3825             2.4333
```

INCREASING ACCURACY

- Regression algorithms
  - Ridge Regression
  - Linear Regression
  - K-Neighbors
  - Decision Tree
  - RANSAC
- Dedicated algorithm for calculating best regressor and best cross validation method
- Cross-Validation
  - Leave-one-out
  - K-fold



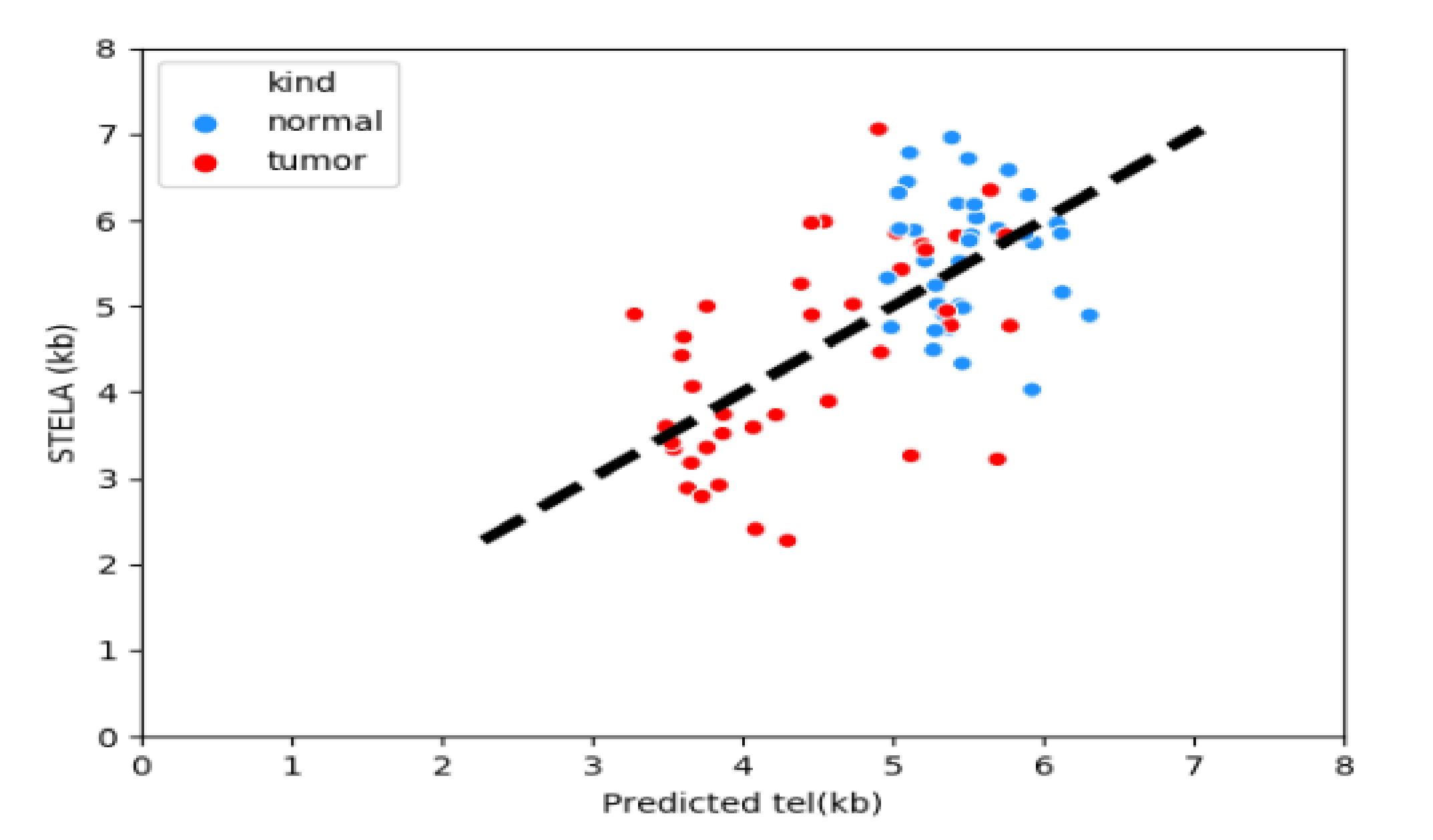
```
# Main script calling the base functions of the pipeline
def main():
    # Define the desired r square value
    target_value = 0.38

    # Get data and create plot
    df = plot_stela_tel()

    # Regression and cross-validation testing
    (r_square, y, predicted, df2, c) = convert_to_features_format(df)

    # Recursive loop until desire target has been achieved
    while r_square < target_value:
        print(r_square)
        main()
```

LEAVE-ONE-OUT



CONSLUSIONS

- Generate a model software to improve the predictions of telomere length in breast cancer samples.
- Aid the prediction accuracy and overall to be useful in generating a breast-cancer specific predictor.
- Chromothripsis seems to have a high weighting for tumour samples. However, this is unsurprising as chromothripsis is only found in tumours.
- Isolation forest improved the output
- A method devoted in finding the best fit with recursion improved the results even more.

FUTURE DIRECTIONS

- A great opportunity to address issues such as the handling of different parameters.
- These pipelines can be used to derive a mean telomere length with relatively better accuracy and at no extra cost.

ACKNOWLEDGEMENTS

Dr Kez Cleal was the supervisor of this case study and this analysis was accomplished under his guidance. The author wishes to recognise this wonderful collaboration and to thank him personally for his advice, useful feedback and insightful comments. The author also wishes to thank his course leaders Dr Kevin Ashelford and Dr Hywel Williams for their detailed informative feedback and suggestions for this study.

REFERENCES

- Cleal, K. et al. 2019. Chromothripsis during telomere crisis is independent of nhej, and consistent with a replicative origin. Genome research. 29(5), pp. 737–749
- Ding, Z. et al. 2014. Estimating telomere length from whole genome sequence data. Nucleic Acids Research 42(9), pp. e75–e75.
- Farmery, J. et al. 2018. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. Scientific Reports 8.
- Gilley, D., Tanaka, H. and Herbert, B.-S. 2005. Telomere dysfunction in aging and cancer. The International Journal of Biochemistry & Cell Biology 37(5), pp. 1000 – 1013. Available at: http: //www.sciencedirect.com/science/article/pii/S1357272504003371. Cancer and Aging at the Crossroads.
- Lai, T.-P. et al. 2017. A method for measuring the distribution of the shortest telomeres in cells and tissues. Nature Communications 8(1), p. 1356. Available at: https://doi.org/10.1038/ s41467-017-01291-z.
- Lewinson, E. 2018. Outlier detection with isolation forest Available at: https:// towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e. [Online; accessed 08-March-2020].
- Liu, F. T., Ting, K. M. and Zhou, Z. 2008. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422.
- Pedregosa, F. et al. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12(null), p.2825–2830.
- Vera, E. and Blasco, M. 2012. Beyond average: Potential for measurement of short telomeres. Aging 4, pp. 379–92.