# Improving every child’s chance in life.

# The Issue: as many as 23% of SA children in the first year of school are developmentally vulnerable

on 1 or more domains (social, emotional, physical, language/cognitive and communication/general knowledge) according to the [Social Health SA Atlas](http://www.publichealth.gov.au/data_online/2012/notes/notes_data_online_aust_2012_option1.pdf).

Early childhood development is a strategic priority for the SA government. It has been established that the first 5 years of a child’s life is the major influence on their future succes.

Our rough estimate of the incurred cost to SA Government is up to $20M per year.

**There is a need to understand how the risk of early childhood developmental issues can be reduced so that every child has every chance in life.**

# What we did:

We gained insights into early childhood development vulnerability and identified the key drivers and the levers that the SA government, community and individuals can apply to improve the situation. We did that by applying the power of cutting-edge data science techniques to the wealth of SA and national data.

We visualised our findings by creating a publically accessible web app that visualises community risk via an indicator to discover if a person lives in a community where their child's development might be at risk and then empowers them to make an informed decision.

# What we found:

After accounting for the population specifics and socioeconomic factors, such as unemployment and Index of Relative Socio-economic Disadvantage, the most important drivers impacting early childhood development were; the lack of a motor vehicle in the household, an inability to afford medication and smoking during pregnancy.

# Intervention strategies we recommend to be implemented by the government:

* To improve access to medical services, provide taxi vouchers or mini buses to health centres on a regular basis, particularly for pregnant women and women with children of early age
* Provide subsidies for medication specific to pregnant women, mothers and children of early age
* Educate community and nurses to deliver anti-smoking programs to women of child-bearing age

# Appendix

## Our approach

We identified the key data sources in the public domain that contained factors that could potentially influence early childhood development ([please see appendix for detailed information](#_Data_sources:) on the data sources used).

The data sources comprised 2000+ fields. To filter though the data fields and identify the key drivers of early childhood development, we applied the following business analysis, statistics and data science techniques:

* Gain business understanding of the data in the data sources. Research to confirm the data field definitions and measurement scales. Exclude from consideration data that was not likely to influence early childhood development (for example, Aged Care data). This reduced the number of data fields to 93.
* Identify and resolve any data quality issues (for example missing values, level of noise). R statistical computing software was used to achieve this.
* Establish and exclude redundant data fields via [advanced correlation analysis](#_Correlation_analysis_to). This reduced the number of data fields to 61. R statistical computing software was used to achieve this.
* Apply advanced data science techniques ([generalised boosting models](#_Boosting_Trees_for), [random forests](#_Random_forests) and [MARS](#_Multivariate_adaptive_regression)) to
  + identify the key factors that influence the percentage of early childhood development vulnerability in an SLA
  + establish which levers and how they can be applied by the Government, community and individuals to reduce the risk of early childhood development impairment.
* Visualise the findings in a Tableau dashboard. To enable effective visualisation of the findings, express the data science results via a linear regression model in R
* **Develop a web app that…**

## Data Notes on Early child development: AEDI, 2009

**Source for all *Early child development* data:** Compiled by PHIDU from AEDI 2009 Research CURF Version 1, Released April 2011, DEEWR

• Developmentally vulnerable on 1 or more domains

• Developmentally vulnerable on 2 or more domains

• Physical health and wellbeing domain - developmentally vulnerable, at risk and on track

• Social competence domain - developmentally vulnerable, at risk and on track

• Emotional maturity domain - developmentally vulnerable, at risk and on track

• Language and cognitive (school based) domain - developmentally vulnerable, at risk and on track

• Communication skills and general knowledge domain - developmentally vulnerable, at risk and on track

**Notes for all *Early child development* data:**

In 2009, the Australian Early Development Index (AEDI), which provides a picture of early childhood development outcomes for Australia, was undertaken nationwide. In the 2009 data collection, information was collected on 261,147 Australian children (97.5 per cent of the estimated five-year-old population) in their first year of full-time school between 1 May and 31 July. A follow-up data collection occurred in some small areas in 2010. In addition, small numbers of children were combined so that more communities could have their results released.

The initial results from the AEDI provide communities and schools with information about how local children have developed by the time they start school across five areas of early childhood development: physical health and wellbeing, social competence, emotional maturity, language and cognitive skills (schools-based), and communication skills and general knowledge.

The AEDI results report on the number of children scoring in the following percentile ranges: 0 to 10th percentile (developmentally vulnerable), 11th to 25th percentile (developmentally at risk), 26th to 50th (on track lower range) and above the 50th percentile (on track higher range).

The data shown include children who were developmentally vulnerable (0 to 10th percentile) in one or more/ two or more domains; children in each domain who were assessed as being developmentally vulnerable (0 to 10th percentile), developmentally at risk (11th to 25th percentile) or developmentally on track (above the 25th percentile).

Data are not shown for areas where there were less than 15 children tested.

## Software used

* R v2.15.3
* Rstudio 096.330
* Tableau Desktop v8
* Tableau server v8
* SQL Server 2012
* Amazon Web Services
* Visual Studio 2012
* Java
* Microsoft Excel 2013

## Data sources:

### 1. Social Health Atlas of Australia: Notes on the Data

Published 2012 (SA)   
Published by Federal Public Health Information Development Unit

[http://www.publichealth.gov.au/templates/blue/images/banner.gif](http://www.publichealth.gov.au/index.php)

The Public Health Information Development Unit (PHIDU), located at [The University of Adelaide](http://www.adelaide.edu.au/), was established in 1999 to assist in the development of public health data, data systems and indicators.

<http://www.publichealth.gov.au/data/>

([data notes](http://www.publichealth.gov.au/data_online/2012/notes/notes_data_online_aust_2012_option1.pdf))

### 2. [Australian Early Development Index Data](http://data.sa.gov.au/dataset/australian-early-development-index/resource/01278622-09c8-40b7-86a9-67b9819d0b25)

**Source:** Dept for Education and Child Development of SA

### 3. Health service locations

**Source**: Child and youth health ([cyh.com](http://cyh.com/))

### 4. Geographic boundary files. Australian Bureau of Statistics (ABS)

**Source:** <http://abs.gov.au/geography>

* Statistical local area (SLA)
* Postcode area (POA)

## Technical methods

### Correlation analysis to establish redundant data fields

The absolute values of pair-wise correlations are considered. If two variables have a correlation of 95% or higher, the method reviews the mean absolute correlation of each variable and removes the variable with the largest mean absolute correlation.

### Boosting Trees for Regression and Classification Introductory Overview

Over the past few years, the computational approach of stochastic gradient boosting has emerged as one of the most powerful methods for predictive data mining. Some implementations of these powerful algorithms allow them to be used for regression as well as classification problems, with continuous and/or categorical predictors.

**References**

[1] Y. Freund and R.E. Schapire (1997). “A decision-theoretic generalization of online

learning and an application to boosting,” Journal of Computer and System

Sciences, 55(1):119-139.

[2] J.H. Friedman (2001). “Greedy Function Approximation: A Gradient Boosting

Machine,” Annals of Statistics 29(5):1189-1232.

[3] J.H. Friedman (2002). “Stochastic Gradient Boosting,” Computational Statistics

and Data Analysis 38(4):367-378.

[4] J.H. Friedman, T. Hastie, R. Tibshirani (2000). “Additive Logistic Regression: a

Statistical View of Boosting,” Annals of Statistics 28(2):337-374.

[5] B. Kriegler and R. Berk (2010). “Small Area Estimation of the Homeless in Los

Angeles, An Application of Cost-Sensitive Stochastic Gradient Boosting,”Annals

of Applied Statistics 4(3):1234-1255.

[6] G. Ridgeway (1999). “The state of boosting,” Computing Science and Statistics

31:172-181.

[7] C. Burges (2010). “From RankNet to LambdaRank to LambdaMART: An Overview”, Microsoft Research Technical Report MSR-TR-2010-82

### Random forests

Random forests are an ensemble learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

**References**

Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.

Ho, Tin Kam (1995). "Random Decision Forest". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests". IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8): 832–844. doi:10.1109/34.709601.

Amit, Yali; Geman, Donald (1997). "Shape quantization and recognition with randomized trees". Neural Computation 9 (7): 1545–1588. doi:10.1162/neco.1997.9.7.1545.

Kleinberg, Eugene (1996). "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition". Annals of Statistics 24 (6): 2319–2349. doi:10.1214/aos/1032181157. MR 1425956.

Dietterich, Thomas (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization". Machine Learning: 139–157.

Criminisi, Antonio; Shotton, Jamie; Konukoglu, Ender (2011). "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning". Foundations and Trends in Computer Vision 7: 81–227. doi:10.1561/0600000035.

Lin, Yi; Jeon, Yongho (2002), "Random forests and adaptive nearest neighbors", Technical Report No. 1055, University of Wisconsin

Liaw, Andy (16 October 2012). "Documentation for R package randomForest". Retrieved 15 March 2013.

Deng,H.; Runger, G.; Tuv, E. (2011). "Bias of importance measures for multi-valued attributes and solutions". Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). pp. 293–300.

Altmann A, Tolosi L, Sander O, Lengauer T (2010). "Permutation importance:a corrected feature importance measure". Bioinformatics. doi:10.1093/bioinformatics/btq134.

Tolosi L, Lengauer T (2011). "Classification with correlated features: unreliability of feature ranking and solutions.". Bioinformatics. doi:10.1093/bioinformatics/btr300.

Deng,H. (2013), Guided Random Forest in the RRF Package, Technical Report

### Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is a form of [regression analysis](http://en.wikipedia.org/wiki/Regression_analysis) introduced by Jerome Friedman in 1991.[[1]](http://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines#cite_note-1) It is a non-parametric regression technique and can be seen as an extension of [linear models](http://en.wikipedia.org/wiki/Linear_model) that automatically models non-linearities and interactions between variables.

**References**

[1] S. Arlot and A. Celisse. A Survey of Cross-Validation Procedures for Model Selection. Statistic Surveys, 2010.

[2] Y. Bengio and Y. Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. J. Mach. Learn. Res., 5:1089U1105, 2004. } http://jmlr.csail.mit.edu/papers/v5/grandvalet04a.html.

[3] Julian Faraway. Extending the Linear Model with R. CRC, 2005. http://www.maths.bath.ac.uk/~jjf23.

[4] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers.Revised version of Technical report HP Laboratories. HP Labs, 2004. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RmS.

[5] Jerome H. Friedman. Fast MARS. Stanford University Department of Statistics, Technical Report 110, 1993. http://www.milbo.users.sonic.net/earth/Friedman-FastMars.pdf, http://www-stat.stanford.edu/research/index.html.

[6] Jerome H. Friedman and Bernard W. Silverman. Flexible Parsimonious Smoothing and Additive Modeling. Technometrics, Vol. 31, No. 1., 1989

[7] Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning (2nd ed.).Springer, 2009. http://www-stat.stanford.edu/~hastie/pub.htm